THE UNIVERSITY
of EDINBURGH

The phylodynamics of infectious diseases of livestock: preparing for the era of large-scale sequencing

Matthew Hall

THE UNIVERSITY
of EDINBURGH

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the University of Edinburgh

2015

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated overleaf, and that this work has not been submitted for any other degree or professional qualification.

Parts of this work have been published in *mBio* and on *arXiv.org*.

Matthew Hall

25 August 2015

Chapters 1 and 7: Parts of these chapters have been accepted for publication in *Revue Scientifique et Technique (OIE)*.

Chapter 2: The previously unpublished SAT 2 isolates were sequenced at the FAO World Reference Laboratory for Foot-and-Mouth Disease and provided by Dr. Nick Knowles and Dr. Jemma Wadsworth. The chapter was published as Hall et al. [59].

Chapter 3: The selection of simulated scenarios in this chapter arose from discussions involving Professor Andrew Rambaut, Professor Andrew Leigh Brown, and members of their groups.

Chapters 5 and 6: An earlier version of these chapters was published in the *arXiv.org* preprint server as Hall and Rambaut [60], and it has subsequently been accepted for publication in *PLOS Computational Biology*.

Chapter 6: Dr. Trevor Bedford provided the idea of using a noninformative sequence for known cases for which no sequence is available.

Abstract

A rapid increase in the amount of available pathogen genetic data, which is ongoing and likely to continue for the foreseeable future, presents new opportunities and challenges in molecular epidemiology, and in the emerging field of "phylodynamics", which seeks to unify the study of the evolutionary and epidemiological dynamics of pathogen populations. This thesis explores some of these challenges and opportunities, with a focus on pathogens infecting livestock and poultry. I conducted analyses of sequences from two serotypes of foot-and-mouth-disease virus (FMDV) in order to investigate the global population dynamics of the virus. For serotype SAT 2, the amount of publicly available genomic data is still small enough that all of it could be included in a single analysis. A particular focus was the origins of historical outbreaks occurring in North Africa and the Middle East, outside the endemic area for the serotype. The results suggested sources for these in countries just south of the Sahara, and that the viruses responsible for three outbreaks occurring in 2012 were the result of separate introductions. For serotype O, including every available sequence was not feasible and the data had to be sub-sampled. Little research has been conducted on how to design a sampling strategy for sequence analysis of pathogens, an issue of increasing importance, so a simulation study was conducted to identify one. This suggested that, when reconstructing the temporal and spatial dynamics of a structured population of pathogens or infected individuals, it is preferable to stratify by subpopulation and by time period. The type O analysis itself showed that the south-east Asian

topotype moves between countries according to cattle trade networks, but that geographic proximity is also important for strains from southern Asia and the Middle East. With genetic data available at an epidemiological resolution that was previously inconceivable, there are opportunities for new types of inference. For example, if we can acquire a sequence from all or most infected cases in an epidemic, they can inform inference of who infected who, complementing traditional contact-tracing approaches. I introduce a novel phylodynamic method for the simultaneous reconstruction of phylogeny and transmission tree for an epidemic in a situation where every infected host or premises can be identified and a sequence acquired from most of them. The performance of this method was demonstrated using simulated data, and then it was applied to reconstruct both trees from the 2003 H7N7 avian influenza outbreak in the Netherlands.

Lay summary

Recent advances in genome sequencing technology mean that, in future, much more genetic data of all sorts will be available to researchers than has been in the past. In particular, a large increase in the number of historical sequences from viruses, bacteria, and other infectious organisms, and the ability to sequence newly isolated strains quickly and cheaply, provides an important new opportunity to epidemiologists. The reconstruction of the history of pathogen strains can give important insights into how they have spread within their host populations in the past. Dealing with this huge source of new data requires new methods and analysis techniques, and this thesis focuses on some of the issues involved. In the first half, two analyses of foot-and-mouth disease virus sequences are conducted in order to investigate the global patterns of spread of this important livestock pathogen. This identifies the origins of three outbreaks taking place in North Africa and the Middle East during 2012, and also suggests that transmission of the virus between countries is the result of the cattle trade and geographic proximity. To prepare for this work, a simulation exercise was performed to identify the best way to pick a collection of sequences for analysis. The second half of the thesis outlines a method to simultaneously reconstruct both the phylogeny, which depicts the ancestral history of the sampled pathogens, and the transmission tree, which depicts which case in the epidemic infected which other, from genetic data collected and sequenced from all or most cases in an epidemic. This method was then applied to both a simulated set of sequences (in order to show that it

can produce accurate reconstructions) and data from a 2003 outbreak of avian influenza in the Netherlands.

Acknowledgements

This PhD project was funded by EPIC, the Scottish Government's centre for expertise on animal disease outbreaks. I also thank the SMBE for the graduate student travel award that allowed me to attend the 2014 meeting.

I thank my supervisors, Mark Woolhouse, Andrew Rambaut, and Ruth Zadoks, for all their support and input in getting me and the contents of this document to this point.

I thank the other researchers in pathogen phylogenetics from Epigroup, the Rambaut group, and the Leigh Brown group for providing an intellectually stimulating environment.

I thank Epigroup and Ashworth as a whole for being one of the nicest place to do a PhD that I can imagine. A special thank you to the boardgamers.

I thank my parents for general support and for proofreading a large document on a subject they are almost entirely unfamiliar with.

A final thank you to Melissa for everything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The short life-cycles and fast mutation rates of viruses and other pathogens are such that evolutionary and epidemiological processes occur on similar time-scales [56, 115]. Thus, events that take place over the course of the infection of a population of hosts and affect the population of pathogens responsible for it may be detectable from sequence data, even on a time-scale as short as that of an individual outbreak or epidemic. This makes genetic analysis of pathogens a useful tool in infectious disease epidemiology.

This field is rapidly coming of age as sequencing technology becomes much faster and cheaper [101] and the amount of available data increases as a result. The number of nucleotide base pairs in all the sequences available in the public NCBI Nucleotide database is increasing by over 20% per year for both viruses and bacteria [12]. As a result, the genetic characteristics of pathogen populations can be sampled at increasingly high resolutions; where before only a handful of sequences might be known for a given infectious agent, limiting the inference that could be drawn from them, now datasets are very large and in some cases even approaching comprehensive on a clinical level. For example, sequences are

available for at least 60% of the UK population of HIV-infected men who have sex with men [92]. The availability of such data on an unprecedented scale opens up wholly new lines of investigation, but also introduces challenges in coping with a vast increase in the amount of available information.

The term "phylodynamics" has been coined to describe the study of how the evolutionary and epidemiological behaviour of pathogens interact with each other [56], but as Kühnert et al. [91] point out, it has come to be applied to two rather distinct ideas. These correspond to the two sides of the divide between theory and data. Some studies investigate the effect of particular evolutionary events or population structures on disease dynamics [53, 88, 163], but this work to date has been largely theoretical and has used simulated data. "Phylodynamic" work on real data, on the other hand, uses phylogenetic methods to describe the relationships between individual pathogen isolates, as well as the timing of their ancestral history, based on a comparison of their DNA or RNA sequences. These relationships and timings are then used to inform inference about epidemiological processes. The effect of such processes on the genome itself is largely not considered; mutations are effectively treated as neutral. Indeed, sequence regions that are known to be under strong selection may be excluded as unsuitable for analysis (see e.g. [96]). This is undoubtedly a simplification, but it does allow for genuine insights to be discovered from real data. Kühnert et al. [91] go as far as to reserve the term "phylodynamics" for the former approach only, using "phylogenetic epidemiology" for the latter. While this terminology does not appear to have caught on as yet, if it is accepted then the work in this thesis comes very much from the phylogenetic epidemiology end of the scale.

1.1 Phylogenetic reconstruction

A phylogeny or phylogenetic tree is a diagram depicting the evolutionary history of a set of organisms. In mathematical terminology, a *graph* is a diagram made up of *nodes* or *vertices* joined by *edges*. A *tree* is a graph that fulfils three additional conditions. The first is that it is *simple*; there is at most one edge joining any two nodes. The second is that it is *connected*; for any two nodes there is a path between them along the edges. The third is that is *has no cycles*; there is no path from a node back to itself that does not re-use an edge. The edges of a tree are often referred to as *branches*. In a phylogenetic tree, each node represents an organism, either one that has been sampled for analysis, or an unsampled common ancestor of those that have been sampled. The edges represent the genetic lineage that leads from ancestor to descendant.

Graphs, and trees, can be *directed*, in which case each edge has an orientation. In phylogenetics this orientation represents the passing of time from past to future, and it is normally established by identifying a *root node* for the tree, which represents the common ancestor of every organism in the sample. (The assumption is made that every such organism did indeed share such a common ancestor.) Then the edges are directed such that every path points away from the root, representing the temporal direction of descent this common ancestor. With this directionality established, every node except the root has a *parent* node, one that is closer to the root in time that itself, representing an ancestor of the organism represented by the node. Some also have *child* nodes, representing descendants. Those that have no children correspond to the contents of the sample; these are called *external nodes* or *tips*. Those with children correspond to unsampled ancestors and are known as *internal nodes*. Although this is not an essential requirement, all trees considered in this thesis are *binary*, in that every internal node has exactly two children. The root node is often depicted as having a branch leading back in time,

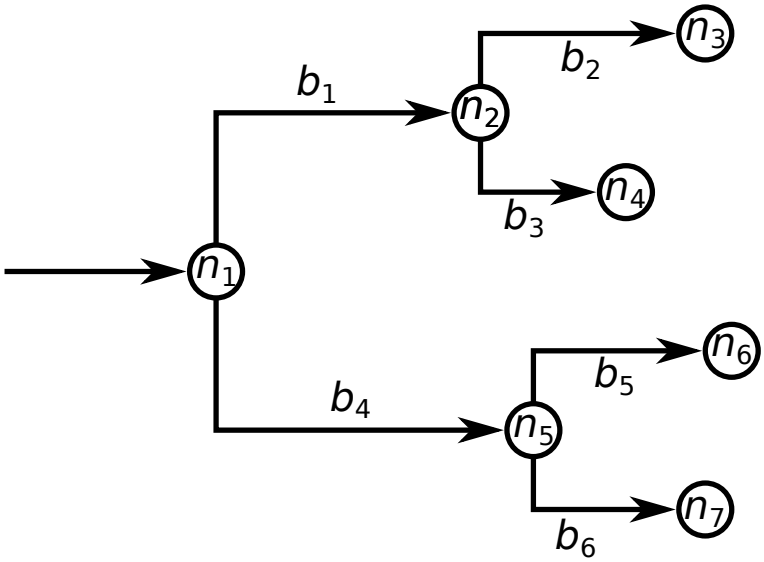


Figure 1.1: An example phylogenetic tree. Nodes are marked n_1 to n_7 and edges b_1 to b_6; arrows represent orientation and n_1 is the root node. The nodes n_3, n_4, n_6 and n_7 are external and would correspond to sampled organisms; the rest are internal and represent unsampled ancestors. As an example, the node n_2 has a parent n_1 and two children n_3 and n_4.

but this is only for clarity and does not lead to a node. See figure 1.1 for a simple example of a phylogeny.

Phylogenies are constructed under the principle that the more similar two organisms are, the more recently they shared a common ancestor. There are many ways to compare organisms for similarity, but in the work outlined here it is always done on the basis of a comparison of their nucleotide sequences. The principle is that the more differences there are between sequences, the more mutation has happened since the respective organisms shared a common ancestor. While the simple genetic distance (the number of differing sites) can be used as a measure of difference between sequences, this does not take into account the nature of the mutation process (such as the possibility of back-mutation, and differing probabilities of occurrence for different mutations) and more sophisticated approaches utilise a nucleotide substitution model (for example [61, 142]). Once a measure by which

distances are defined has been picked, a computer algorithm is used to produce a phylogeny, or set of probable phylogenies.

Branches in a phylogeny have lengths, and these lengths usually represent one of two things: the amount of mutation that occurred between the organism represented by the parent node and the organism represented by the child, or the calendar time that elapsed between the existence of the former and the latter. For trees of the latter type, often known as "time trees", information on the time at which the samples represented by the tips were taken is required. For the purposes of phylodynamics, the ability to infer phylogenies on a calendar timescale is generally essential, as this is the only way they can be used to examine epidemiological trends, so time trees are quite standard.

There are a large number of different algorithms of varying complexity used for the purpose of phylogeny reconstruction. Some, such as neighbour-joining, simply attempt to produce a single tree that minimises the amount of evolution that would have to have taken place over the ancestry of the set of samples; others take a statistical approach and fit a model of mutation to the data. The statistical approaches taken can be both frequentist and Bayesian. However, for time trees the Bayesian versions are most common. The approach used in this thesis is that implemented in the program BEAST [39]. Unlike non-statistical or frequentist methods, Bayesian algorithms do not attempt to construct a single "most probable" phylogeny representing the ancestry of the sequences in a sample. Instead they use Bayes' Theorem to investigate the posterior probability distribution of phylogenies. Specifically, if T is a tree and S a set of sequence data:

$$p(T|S) = \frac{p(S|T)p(T)}{p(S)}$$

The term $p(T)$ is the *prior* probability; it is a probability calculated from a

distribution for the set of possible trees that we give before the analysis is conducted (i.e. before the sequence data is known). The term $p(S)$ is a normalising constant, and if a formula for the *likelihood* $p(S|T)$ is available then the *posterior* probability $p(T|S)$ can be calculated at least up to multiplication by this constant. If this probability can be expressed analytically, the probability of any tree can be calculated. In practice, however, these probability spaces are so complex that exhaustively doing this for every possible tree is not feasible; instead Markov Chain Monte Carlo (MCMC) sampling is used to obtain a representative sample from the posterior distribution, usually consisting of thousands of trees.

The tree is often not the only variable of interest from a statistical phylogenetics analysis; indeed in some cases it is not even the primary variable of interest. These procedures work by fitting a model to the genetic data, and we may be primarily interested in the parameters of that model. The MCMC process also allows for simultaneous estimation of such parameters. An example of how this works in recovering epidemiological model parameters can be found in section 1.1.1, but rates of nucleotide substitution are also generally estimated in this way, and the phylogeographic methods outlined in section 1.1.2 also work by the same principles.

1.1.1 The tree prior

Suppose ϕ is a collection of parameter values for a mathematical description of the behaviour of a population of pathogens, which we are interested in estimating for epidemiological purposes. For example, the basic reproduction number R_0 may be one of the parameters, or it may be possible to derive it from them. Suppose we also have sequence data S and a phylogenetic tree T describing the ancestral relationships between the sequences. Again by Bayes' theorem we have

the following equation for the posterior probability of both ϕ and T given S:

$$p(T, \phi|S) = \frac{p(S|T, \phi)p(T, \phi)}{p(S)}$$

It is assumed that S is conditionally independent of ϕ given T [138]; in other words that since ϕ are the parameters that generate the tree, they provide no further information if the tree itself is known. Since $p(T, \phi) = p(T|\phi)p(\phi)$, this then becomes:

$$p(T, \phi|S) = \frac{p(S|T)p(T|\phi)p(\phi)}{p(S)}$$

This is the joint probability distribution for the values of ϕ and T. It provides a method for estimating the values of ϕ. $p(\phi)$ is determined by a prior, and it remains to formulate an expression for $p(T|\phi)$, the probability of the tree given the parameter values. With this available, MCMC can be used to sample from $p(T, \phi|S)$ and obtain an estimate of the marginal probability distribution $p(\phi|S)$. Other properties, such as nucleotide substitution rates, can be estimated at the same time.

For the particular model whose parameter values are ϕ, there are several alternatives currently available. Most commonly it consists of the parameters of a coalescent process; this is a method from population genetics that uses the rate at which lineages would be expected to split over time to estimate node times of the phylogeny [57], if they were part of an idealised, freely-mixing population of a certain size. This size is known as the *effective population size*, N_e. Because genuine populations are much more complicated than the idealised cases used in the models, the relationship between N_e and the census population size is very difficult to quantify and the actual numerical values of the former are rarely interpreted, but changes in it are still assumed to reflect changes in the former. In fact, the nature of the coalescent model is such that N_e is confounded with the

generation time τ of the organism, so what is truly estimated is the product $N_e\tau$. $N_e\tau$ may be assumed to be obeying simple mathematical rules, such as constant size, or exponential or logistic growth [114, 116]. Estimates of the parameters of these growth models are estimated by the MCMC and if it can be assumed that the pathogen population is experiencing the appropriate kind of growth, epidemiological parameters such as reproductive numbers can be obtained from these [114, 162].

As an alternative to these simple models for the behaviour of the effective population size, methods in the Bayesian skyline [36] family break up the history of the genealogy into a finite number of disjoint intervals and allow $N_e\tau$ to take a different constant value on each of them. A different step function for $N_e\tau$ over time is produced for each MCMC sample, and these can be smoothed by averaging over an entire sample set to give a reconstruction of the behaviour of the population size. These methods can give a good idea of the changes that the population size has undergone since all samples had a common ancestor, but direct estimation of important parameters from them is not straightforward. This family includes the original Bayesian skyline plot [36], and its successors the skyride [102] and skygrid [52], which introduced refinements to the model.

Use of the coalescent process makes the assumption that the set of samples represents a very small proportion of the total population of organisms. If an epidemic is well-sampled, and if within-host genetic variation is assumed to be negligible, this may well not be true [139]. Additionally, as mentioned above, it does not readily provide a means to directly estimate epidemiological parameters such as reproduction numbers unless the behaviour of the pathogen is relatively simple. The alternative is to assume that the tree is generated by an explicit epidemiological model, such as one of the standard compartmental models (SI, SIR, etc.) which divide the population of hosts up into categories and posit rates of movement between them, whose actual parameters are those of interest. Examples

are the birth-death model of Stadler et al. [139] and the SIR coalescent of Volz et al. [155]. Rasmussen et al. [117] also provide a general framework for integrating time series and geneological data using a stochastic compartmental model.

1.1.2 Phylogeography

Phylogeography concerns the use of phylogenetics to study genetic variation in a geographical context. The combination of the relationships between organisms revealed by a phylogenetic tree and the locations at which the organisms were sampled, taken together, provide information about the movements made by ancestral organisms. In the context of infectious diseases, where phylogenetics and phylodynamics can be used to study the temporal dynamics of the organism [68], a phylogeographical analysis can help to reconstruct the movement of the pathogen between locations over time. Within a Bayesian MCMC framework, movements between locations are estimated simultaneously with the phylogeny and other parameters of interest. Two models of location are possible: discrete and continuous.

The discrete model [94] allows for a finite number of discrete locations (for example, countries); a significant limitation is that only the locations from which the samples were taken are available for the analysis, making the sampling scheme important. The results of the analysis can give statistic tests using Bayes factors (BFs) for the hypothesis that the rates of movement of the pathogen between two particular locations is nonzero (in each direction, if necessary), allowing maps to be drawn of the linkages between locations that are well-supported. Alternatively, the Markov Jumps procedure [103] can be used to reconstruct, for each MCMC sample, a realisation of the stochastic process of lineage movement between locations in the ancestry of the set of sequences. While geography is probably the most commonly-used application of this discrete model, it can also be applied to any other discrete

characteristic of the set of samples. For example, in pathogens infecting multiple species it has been used to reconstruct transitions between types of host.

The continuous model [95] requires each sample to have a location on a continuous scale (such as latitude and longitude), assumes that lineages move through the landscape in a process of diffusion, and infer the likely locations of the common ancestors. A model similar to that of the relaxed molecular clock allows rates of diffusion to be different along different branches. This has the advantage that it does allow ancestral location states to be unsampled in the data, but the disadvantage is that precise coordinates for each isolate are required, and the method at present does not ensure that the inferred ancestral location actually makes sense. (For example, if two samples were on opposite sides of a sea, the common ancestor might well be inferred as being present in the middle of it.)

A limitation of both these models for geography in BEAST is that they infer a diffusion process between locations which are treated as characteristics of lineages that are wholly separate to the coalescent population model used to reconstruct temporal dynamics. In actual fact, one would expect geography to influence the population such that it was not freely-mixing. Coalescent-based methods of this sort do exist and are known as *structured*, although the mathematics of this is considerably more complicated than the single-population model. The overall population is divided into subpopulations known as *demes* and lineages are assumed to migrate between demes at rates that are parameters of the model. A recent paper by De Maio et al. [30] introduced a fast approximation for structured coalescent inference in BEAST; this can be used to replace the discrete phylogeography model. It does, however, at present assume that the size $N_e\tau$ of each deme is constant over time.

1.2 Transmission tree reconstruction

Perhaps the most ambitious potential application of molecular sequence data to the epidemiology of infectious diseases would be to use it to reconstruct the transmission history of the disease, providing the links of direct spread amongst a population of infected individuals or premises. The ultimate target of such investigations is the recovery of the transmission tree of the epidemic, a diagram of who infected whom for all hosts that experience an infection, sometimes combined with information on the time that each became, and ceased to be, infected or infectious. Traditional methods for investigation of the transmission tree have relied upon contact tracing, a labour-intensive procedure that must deal with many unknowns. Genetic data now offers a promising new source of information. If the rate of mutation is sufficiently fast, genome sequences for viruses, bacteria or other infectious agents taken from different hosts will be distinct from each other. The positive relationship between the similarity of two sequences taken from pathogen isolates and the closeness of the ancestral relationship between those isolates can be extended to the relationship between the hosts that the isolates came from: a close relationship between pathogens implies that the hosts were close to each other in the transmission tree. This principle opens up the possibility that the tree can be reconstructed using a new type of data that was previously invisible to the naked eye, so long as isolates can be acquired from enough hosts (and, if this is to be conducted while the outbreak is ongoing, quickly enough) to make inference useful. Traditional epidemiological data from contact-tracing or other sources could also be used to augment the procedure.

Ideally, samples would be taken from every host, a natural prerequisite being that all hosts can be identified in the first place. This is more likely for some pathogens, and some host populations, than others; promising situations are those in which all potential hosts will be closely monitored. This is one reason why

work on this topic has often been undertaken on outbreaks occurring in farmed animals. The "host", the infected unit, is taken to be a farm rather than an individual animal, as it is generally of more interest to determine which farms infected which others than how the disease spread from animal to animal. As considerable resources will often be expended to stamp out the disease, at least in high-income countries, identification of all infected farms is quite likely. Work has been published reconstructing the tree for outbreaks of foot-and-mouth disease virus [26, 27, 105, 173], avian influenza [171], and salmon infectious anaemia virus [3]. However, as perfect sampling is unlikely in most circumstances, many of the most up-to-date methods are appropriate for imperfect and even quite sparse isolate collection [104, 107]. The motivation for such work is often to design procedures to reconstruct the tree for endemic disease, but they are also appropriate for poorly-sampled outbreaks. Nevertheless, it will always be true that the transmission tree will be only very partially revealed if a small fraction of the population of hosts provides any data.

The power of these procedures should not be overstated. Perfect reconstruction of the transmission tree using genetic data alone would be possible only if pathogen mutation rates were much faster than they actually are; in practice the genetic diversity that accumulates over the relatively short timescale of an outbreak is limited, some isolates taken from different hosts may be found to have identical sequences, and uncertainty regarding transmission routes will never be entirely eliminated. The output of more sophisticated methods will assign a score to inferred links in the transmission tree designating how well-supported the relationship between the hosts is by the data. Due to the lack of resolution that is frequently seen when inference uses genetic data alone, authors regularly stress the importance of including data from traditional epidemiological investigations and prior knowledge about the pathogens and hosts involved in an analysis [35, 76]. Geographical data or estimates of dates of infection can be used to improve the reconstruction, or

contact tracing can be used to rule out some transmission trees. The emergence of a new way to infer pathogen spread should not be taken as a reason to entirely abandon all the old ones.

As with estimation of the phylogeny, the fundamental principle here is that the more similar the genetic sequences for two pathogens are, the more recently they shared a common ancestor, which must have been present in a single host. Care should be taken in situations where the similarity between sequences in fact cannot be taken simply as a proxy for the closeness of the ancestral relationship between the corresponding isolates. There are two major causes for concern. The first is situations of reassortment or recombination, where two pathogens may have a closer ancestral relationship in some parts of their genomes than in others. In an outbreak situation, and presuming that, even if more than one genetic variant is introduced to a host upon infection, the difference between them is not large, this is only likely to be a serious problem in cases of superinfection; if recombination or reassortment takes place within a host, all the resulting variants are still descendants of the strain that caused the infection and have the same ancestral relationship to it, even if they have exchanged genetic material with each other. If, on the other hand, a host is infected twice by quite divergent strains, mixing of genetic information could have a seriously distorting effect on the picture. It is recommended that datasets be checked for recombination or reassortment using a tool developed for this purpose [112], though no approaches have yet been proposed if it is found. A starting point might be to conduct separate analyses of the parts of the sequence on either side of any identified breakpoint.

The second concerning situation revolves around convergent evolution. While the assumption in methods of this type is that mutation is a neutral process, it frequently is not, and some variations may be selected for. If this is so, then genetic similarity between isolates at some sites may not be the result of a close historical relationship, but of the similar environments that they find themselves

in. Software exists to identify such positions in an alignment [113], and if this is suspected for certain sites, those should simply be excluded from the analysis.

The problem of reconstructing a transmission tree given a measure of the genetic distance between two sequences is in fact closely related to the problem of reconstructing a phylogeny, and similar approaches have been used: simpler ones attempt to find the single tree which keeps the amount of mutation required to a minimum, whereas the more complex construct an ancestry by fitting models of transmission and mutation to the sequence data and include some measure of uncertainty in the output. The phylogeny itself is of relevance, because internal nodes in it correspond to points at which a lineage was present in one host and subsequently split; if descendant nodes are sampled from more than one host, at least one transmission is implied.

There are broadly three classes of transmission tree reconstruction method, of increasing complexity. The simplest assume that a sampled sequence is entirely representative of the strain which infected the corresponding host over the full period of its infection. The intermediate group still assume that each host was infected by one lineage, but allow for mutation of that lineage; any sequence is taken to be entirely representative of the pathogen population in the host at the time of sampling. The most complex class acknowledge that multiple, genetically distinct, lineages can co-exist within a host at the same time. Figure 1.2 provides an illustration of the three approaches. The most complex model is not necessarily the most appropriate to the problem; the assumptions made in the simpler versions have enabled recent work on the detection of unsampled cases, and more basic models may also be preferred for reasons of computational time.

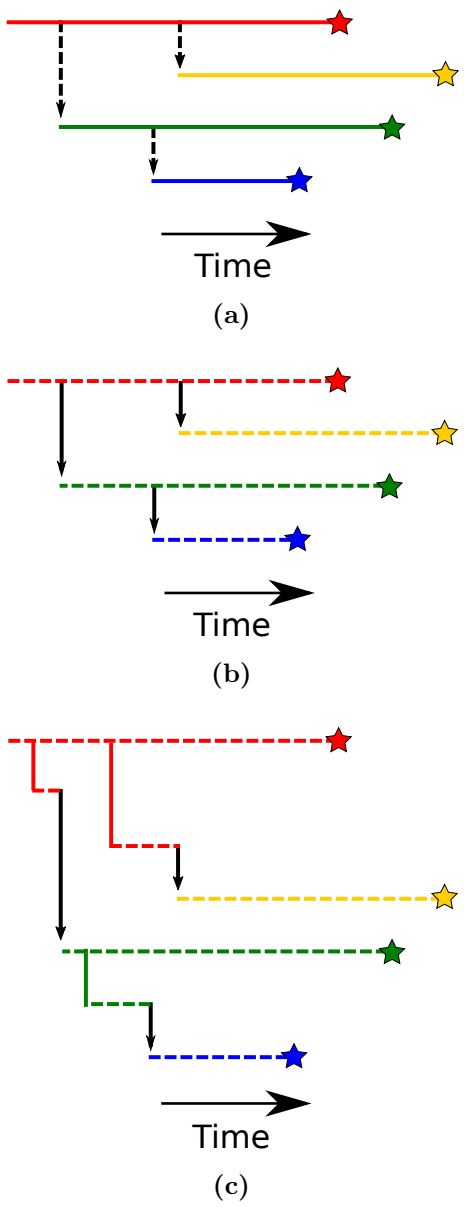


Figure 1.2: Illustration of the three basic approaches to transmission tree reconstruction using genetic data. Stars represent the sampling of isolates from hosts; each horizontal line is a distinct pathogen lineage and is coloured by the host it is present in. Black vertical arrows represent transmissions between hosts, and dashed lines are undergoing mutation. (a) Mutation is a consequence of transmission and only one lineage is present in each host. (b) Mutation occurs within-host but only one lineage is present in each. (c) multiple lineages per host.

1.2.1 Within-host genetic uniformity

The most rudimentary way to infer a transmission network from a set of genetic isolates is to construct a tree that minimises the total genetic distance between them, under the assumption that as few mutations as possible were responsible for the observed sequences [136]. Each sequence is taken to be uniquely representative of the pathogen strain infecting each host, and the transmission process is not modelled in any way. This tree is in a mathematical concept known as the minimal spanning tree, and it has similarities to minimum evolution methods for phylogeny reconstruction. However, it is not identical, because phylogenetics reconstructs a tree with sequences assigned only to leaf nodes, whereas every node in the minimal spanning tree corresponds to a sequence. This approach has the advantage of simplicity; as no assumption of direct transmission is made, links in the network can correspond to any number of intervening hosts and, in fact, this approach is often used to infer transmission histories between epidemiologically unrelated samples [18, 73]. However, it has many inadequacies [124]. It outputs only a single transmission tree, even if large numbers fit the distance matrix equally well, and gives no indication of whether particular ancestral relationships are highly supported by the data or more likely to be spurious. There is also no temporal component to the analysis; the direction along the tree that the pathogens travelled can be at best inferred post-hoc using data about the order of infection, with no guarantee that this approach will be consistent between every pair of isolates.

To deal with the issue of uncertainty, a bootstrapping procedure to overcome the first of these limitations was proposed by Salipante and Hall [124]. A procedure to find the transmission tree that minimises genetic distance while maintaining the order in which sequences were sampled is the *SeqTrack* algorithm developed by Jombart et al. [75]; this also introduces epidemiological data (such as spatial

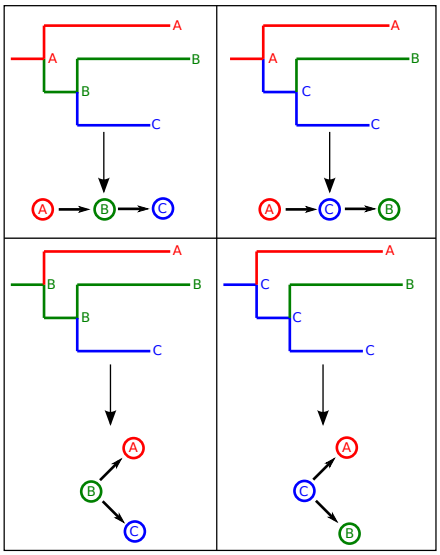
locations) as a means to discriminate between ancestries that are equally likely according to genetic distances.

The *SeqTrack* approach can be improved to accommodate uncertainty by, instead of searching out the single "best" transmission tree, using a Bayesian MCMC procedure to sample from the probability distribution of transmission trees, given the sequences and potentially also epidemiological data (such as spatial locations). As with MCMC estimation of phylogenies, the output is not one but many, potentially thousands, of transmission trees; this set can then be analysed to identify likely pathways of infection. Ypma et al. [171], applied such a procedure to data from the 2003 H7N7 avian influenza outbreak in the Netherlands, incorporating a spatial component defined by a transmission kernel function. The effective assumption, when within-host genetic diversity is ignored, is that mutation is a consequence of transmission. The mutation rate will be expressed in units of mutations per generation, rather than the more common mutations per unit time. While this is certainly a simplification, it can be a useful one; for example it allows for quantification of the number of unsampled links in the transmission chain if the distribution of the serial interval of the infection (the time between successive infections in the chain) is known. If it is likely that two hosts are adjacent in the transmission tree of known hosts, but the number of mutations between them is larger than expected for a single transmission, it would suggest the presence of an unsampled intervening host. This is the principle by which the *outbreaker* algorithm by Jombart et al. [76], another Bayesian MCMC method, can estimate the number of unsampled cases in the transmission chain between the cases from which sequences have been obtained. It also includes a procedure to identify situations where there is likely to be more than one independent introduction to the population of hosts.

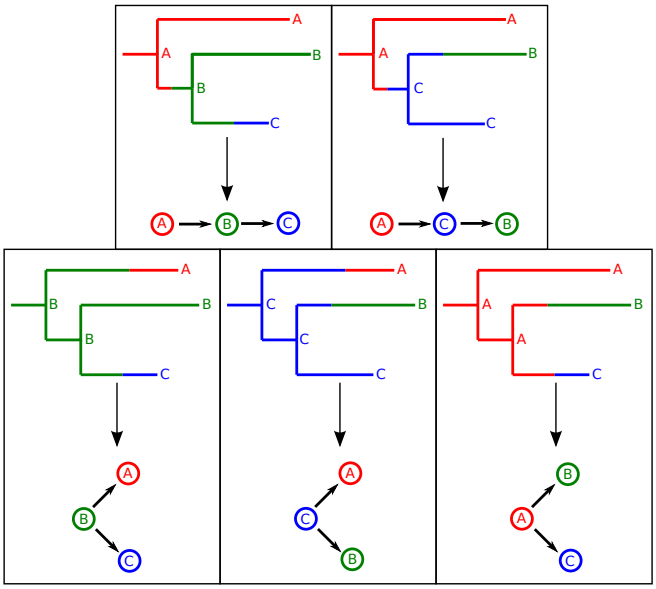
1.2.2 Within-host mutation

If mutations are assumed to occur over the lifetime of a pathogen's presence in a host, but no two genetic variants are allowed to occupy the same host at the same time, the implicit assumption is that lineages split only at transmission. This is a simplification, but is unlikely to be a major one if few mutations are expected to be observed during a host's infection, or if the rate backwards in time at which two lineages "coalesce" to a common ancestor is much faster than the transmission rate of the pathogen between hosts [154]. If one draws a phylogeny, an internal node represents an infection of one host by another, in addition to a common ancestor of pathogen isolates. The work of Cottam et al. [26] explored this by mapping possible transmission histories onto a pre-generated phylogeny for the 2001 UK foot and mouth disease virus (FMDV) epidemic. This is illustrated in figure 1.3a; if we assume that each host was sampled, then each internal node in the phylogeny corresponds to an ancestor of the samples that was present in one of these hosts and by exploring different assignments of nodes to hosts we are in fact exploring different transmission trees. Each internal node must be assigned to the same host as one of its child nodes; a branch whose terminal node is assigned to a particular host corresponds to a lineage existing solely in that host. Cottam et al. then calculated the probability of each possible assignment of these nodes based on epidemiological information about the location of the host farms and their probable infection dates.

The Cottam et al. approach had the limitation that it took a fixed phylogeny as input, and as a result genetic uncertainty was not taken into account. Their dataset was also sufficiently small that they could do calculations by exhaustively assigning the internal nodes of the phylogeny to every possible configuration of hosts. For larger datasets this would prove prohibitive in terms of computational time. The latter limitation can be overcome by use of Bayesian MCMC, which



(a)



(b)

Figure 1.3: Examples of the annotation of the internal nodes of a phylogeny and the correspondence to transmission trees, if one sequence is taken per host in an outbreak amongst three cases. In a), internal nodes represent transmissions, but in b) they do not.

provides a representative sample from the probability distribution of transmission trees without the need to examine every single one. This is the approach taken by Morelli et al. [105], whose method was also the first of this type to not employ an underlying fixed phylogeny. As with Cottam et al., they were working on the 2001 FMDV outbreak and were able to include farm locations in the analysis. The work was extended by Mollentze et al. [104], working instead on rabies samples from South Africa; this second paper extended the procedure to a situation of less consistent sampling by, as with *outbreaker*, allowing for multiple introductions to a study population and for the path of infection between two sampled individuals to pass through unsampled ones, although unlike *outbreaker* the procedure only indicates the presence of such indirect infections and does not enumerate them.

1.2.3 Within-host diversity

Usually, methods allowing for within-host diversity have assumed that only a single genetic variant is passed from one host to another during transmission (in other words, that transmission is a complete bottleneck), but that this single variant is then the source of a large, freely-mutating population. If one were to consider the ancestry of the pathogens within this population that are sampled and sequenced, or are subsequently transmitted to other hosts, it can be represented as a phylogenetic tree. The time of most recent common ancestor of all these sampled or transmitted pathogens is any time after the infection of the host. Each host in the outbreak has such a within-host phylogeny and, if each of these small trees is joined up according to the transmission tree, the result is once again a single phylogeny tracing the ancestry of the samples taken from the entire event. However, no longer is there a temporal correspondence between internal nodes and transmission events.

The methods of the previous two sections have required that two processes be

modelled: the spread of the pathogen between hosts, and mutation. If within-host diversity is to be considered then a model may be required for a third process, which is that occurring within each host. If the "host" is an organism, this will be a model of the dynamics of the population of pathogens infecting it; if it is instead a location, it can instead be a model of the infection as it spreads through the organisms present. In either case, all approaches to date have employed a coalescent process as this model of within-host dynamics, with the population assuming to be freely mixing and its size changing according to a deterministic function. This function may assume an invariant population size [35], or that it obeys exponential growth [173], or that it grows to a peak and then declines [173].

A great advantage of allowing for within-host genetic diversity is that this makes it easy for an analysis to include more than one distinct sequence taken from the same host. A method that assumes that all isolates taken from the same individual or location at the same time are identical obviously cannot deal with data that contradicts this. This is a useful enhancement, as it has been shown in simulation studies that the acquisition of multiple sequences per host can greatly improve the accuracy of inference of the transmission tree [168].

As in the previous two categories, most methods of this type utilise Bayesian MCMC. The first was developed by Ypma et al. [173], who treated every within-host phylogeny as a separate entity. An alternative approach, introduced by Didelot et al. [35] is to modify Cottam et al.'s procedure of annotating the nodes of a single phylogeny with host information. Since internal nodes no longer represent transmissions, a modification must be made; a node must be assigned to the same host as at least one of three nodes: its two children and its parent (see figure 1.3b). This allows for situations in which a lineage in a given host was not the ancestor of any isolate sampled from that host, which is essential in a framework with within-host diversity; e.g. in figure 1.3b, in the bottom right, the common ancestor of the lineages sampled from hosts B and C was actually

present in host A, but is not the ancestor of the lineage sampled from A. The node annotation procedure is convenient because it is highly compatible with existing methods for phylogenetic reconstruction; trees need merely to be annotated with assignments of internal nodes to hosts and infection dates. Didelot et al. applied this only to a fixed overall phylogeny, without accommodating genetic uncertainty.

A radically different framework, which eschews Bayesian MCMC in favour of an importance sampling approach with similarities to approximate Bayesian computation, was recently published by Numminen et al. [107], and avoids modelling within-host dynamics at all, instead simulating a representative set of transmission trees and isolate times of recent common ancestor (TMRCAs), generated by models of transmission and mutation, that conform to a fixed phylogenetic structure. The key advantage of the approach is that it relies on an explicit model of the sampling process, and is therefore of use in situations where sampling is extremely sparse.

1.2.4 Pairwise methods

Some methods eschew any attempt to reconstruct the full transmission tree and instead concentrate on, given any two sequences, attempting to infer the probability that one was the infector of the other. In situations of sparse sampling, this may be the only useful inference that can be drawn in any case. Volz and Frost [154] take this approach, assuming that internal phylogeny nodes correspond to transmissions, and then outlining a method that uses the phylogeny to estimate probabilities of direct transmissions between sampled hosts in a very general framework allowing for complex disease dynamics. Worby et al. [167], while requiring complete sampling, is the first method to incorporate within-host genetic diversity while using a coalescent process for the within-host population which does not assume that transmission is a complete bottleneck, allowing for the

transmission of multiple genetic variants at the same time. Basing inference entirely on pairwise genetic distance, it is also much less computationally intensive than many of the MCMC approaches outlined above. A similarly fast method was presented by Famulare and Hu [44], who identify likely direct transmissions by using a likelihood ratio test of the hypothesis of the time of common ancestor between sequences taken from each case being equal to the sampling date of the earlier one (implicitly assuming no within-host mutation). Where this procedure suggests several potential infectors for a case, a pruning algorithm can be employed to pick a single one, based on, for example, the pair that minimises the time between sampling.

1.2.5 Other approaches

Some investigations have used genetic data as a means to augment traditional contact tracing procedures, without using a combined methodology incorporating both sequences and traditional epidemiological data. For example, Gardy et al. [50] investigated a *Mycobacterium tuberculosis* outbreak using contact tracing and subsequently showed that whole-genome genetic analysis could be used to improve the inference by ruling out connections between cases who were epidemiologically linked but whose pathogen strains, when sequenced, proved to be only distantly related.

An unusual approach was taken by Aldrin et al. [3], who eschewed phylogenetic reconstruction or a model of mutation of any kind entirely, and instead treated the genetic distance between isolates in the same way as geographical distance between locations is treated in spatial models of disease transmission. The probability that one host infected another declines as the genetic difference between their respective sequences increases, according to a transmission kernel function. This was, in fact, combined with a geographical transmission kernel to calculate the probability of

transmission across two landscapes, geographical and genetic. With the parameters of the kernels fit using a maximum-likelihood approach, the probability of each transmission route can be calculated.

1.3 Project aims

The purpose of this PhD project is to explore some of the implications and challenges that the new era of cheap and fast sequencing brings to pathogen phylogenetics and phylodynamics. There are two foci, at very different scales: global phylogeography and transmission tree reconstruction.

The pathogens whose molecular epidemiology is studied here are those affecting livestock and poultry. The reaction of veterinary authorities to outbreaks of these pathogens is rather different to that of medical authorities to human disease, and this has some implications for the kinds of datasets that are likely to be available. An outbreak in a developed country of a disease of significant economic impact, such as foot-and-mouth disease virus (FMDV) or avian influenza, will be regarded as an emergency, and significant resources will be expended to identify infections and bring the event to an end. Identification of every infected animal will be somewhat irrelevant; once an infection is detected in an agricultural facility, the culling of every animal present is likely to follow. Infections are more usefully seen at a farm, rather than animal, level, and this means that the identification of every "infection" is actually quite likely. As alluded to above, the prospects for full transmission tree reconstruction are therefore excellent.

At the other end of the scale, in many parts of the world monitoring is very patchy indeed and the exact epidemiology of the disease is not fully understood. Animals, after all, do not report their illnesses and news of cases may not reach data collectors in resource-poor settings or may be suppressed for commercial

reasons. This is true of FMDV in the areas in which it is endemic. While an increasing amount of sequence data does exist, it has been sampled very erratically over decades. A common approach in the past has been to effectively use every available sequence in an analysis, but this is becoming an outdated practice for two reasons. Firstly, all but the most rudimentary or approximate forms of analysis can require weeks or even months to complete on a dataset of thousands of sequences. Secondly, the availability of a large and diverse collection of sequence data raises questions about how sequences should be selected for analysis in order to avoid bias, or indeed how future studies should be designed.

The next three chapters of this work deal with phylogeography of FMDV in particular, but following a preliminary example in chapter 2, I attempt to place the work on a sounder footing regarding sample selection than has generally been present in previous work, with a simulation study in chapter 3 whose conclusions are applied to real data in chapter 4. Chapters 5 and and 6 go to the other end of the scale and introduce a novel, flexible method to reconstruct simultaneously both phylogenies and transmission trees.

Chapter 2

Reconstructing geographical movements and host species transitions of foot-and-mouth disease virus serotype SAT 2

(This chapter was published in 2013 as Hall et al. [59].)

2.1 Introduction

This chapter provides an initial example of how modern phylogenetics methods can be employed to explore the historical population dynamics of a pathogen, in this case one of the serotypes of foot-and-mouth-disease virus (FMDV).

Foot-and-mouth disease (FMD) is a highly contagious disease of cloven-hoofed

mammals, caused by FMDV, a single-stranded RNA virus of the family Picornaviridae. Seven serotypes exist, two of which (O and A) are found almost worldwide. Another, type C, is more geographically restricted and has not been detected anywhere in the world since 2004, while the Asia-1 serotype is normally confined to southern Asia [87, 128]. The remaining three serotypes are the three Southern African Territories (SAT) viruses, designated SAT 1, SAT 2 and SAT 3, the first two of which are endemic in countries of Africa south of the Sahara; outbreaks due to SAT 3 in domesticated livestock have been confined to a handful of countries in southern Africa. SAT 2, the focus of this chapter, is the SAT serotype most often recorded in domestic animals [144], and is widely distributed across the continent, having been identified as far west as Senegal, east as Ethiopia and south as South Africa. It is further subclassified into fourteen topotypes, designated I to XIV, defined as having 80% nucleotide identity in the VP1 coding region [5, 58].

SAT 2 has crossed the Sahara and caused outbreaks in North Africa and the Middle East on several occasions in recent years. Middle Eastern outbreaks occurred in North Yemen in 1990 [2] and in Saudi Arabia and Kuwait in 2000 [87]. In North Africa, it appeared in Libya in 2003 after an apparent absence from the region for around 50 years [146]. In 2012 outbreaks occurred in Egypt, the Palestinian Territories, Libya and Bahrain [2]. While it might be surmised that the occurrence of so many events in a single year were the result of a single introduction that spread further once established north of the Sahara, Ahmed et al. [2] conducted a genetic study of the viruses involved and found that this did not appear to be the case. Although the bulk of the Egyptian and Palestinian isolates are closely related, those from Libya and Bahrain are of quite distinct lineages. The Bahraini virus is even of a different topotype. Furthermore, one of the samples obtained from Egypt proved to be yet another lineage, distinct from the others collected in the country during the epidemic. For the virus to escape from sub-Saharan Africa four times in one year is unprecedented, and it has been suggested

that the political changes in the region from 2011 onwards (the "Arab Spring") may have played a role (`http://www.bbsrc.ac.uk/news/food-security/2012/120613-f-arab-spring-spread-of-animal-disease.aspx`), as people and their animals were displaced by conflict, or changing governments created new trading relationships and thus new pathways for pathogens to follow. For example, Kandeil et al. [78] note that cattle imports to Egypt from other countries in the Nile basin increased following the Egyptian revolution of 2011, due to improved political relationships between the governments involved.

The epidemiology of the SAT serotypes in sub-Saharan Africa is distinct from that for other serotypes in Africa and elsewhere in that there exists a wildlife reservoir in the form of African buffalo (*Syncerus caffer*) in areas where that species is present [156]. The disease is very rarely symptomatic in buffalo and animals can be persistently infected for a period of several years. As eradication of all infected hosts is therefore not feasible, control has focussed on vaccination and prevention of mixing between buffalo and livestock by means of fencing [71, 156]. Where SAT serotype epidemics have occurred in areas in proximity to areas with buffalo populations they have sometimes been linked to compromised fences [158]. As other wild mammals, such as impala (*Aepyceros melampus*) and other antelopes, are susceptible to FMDV, another cause for concern is the ability of these species to jump over fences and spread infection in that way [156].

It has been some time since the last published phylogenetic analyses of all known RNA sequences for SAT 2 [9]. Since then, the number of available sequences has almost quadrupled, and information on viruses from a much wider range of locations has been added to nucleotide databases. Reclassification of SAT 2 topotypes has also occurred during that time [5, 58]. As summarised in chapter 2, recently-developed phylogenetic techniques enable analyses such as estimation of change in viral genetic diversity over time [36, 102], and the enumeration of historical changes of discrete character states, such as country of origin or host

species, on the phylogenetic tree [94]. This chapter aims to update the complete picture of SAT 2 phylogenetics to include ten years' worth of new sequence data, and apply newer methods in order to examine the source of all recorded outbreaks occurring beyond sub-Saharan Africa since 1990, as well as movement patterns of lineages between countries where the virus is endemic and between host species.

2.2 Methods

2.2.1 The data

Data used were all GenBank records for FMDV serotype SAT 2 that included at least 90% of the VP1 gene (as of May 2013) and sequences for a total of 49 previously unsequenced cell culture grown type SAT 2 FMDVs obtained from the World Reference Laboratory for Foot-and-Mouth Disease Reference Collection. RNA extraction, RT-PCR of the VP1 region and DNA sequencing of these was performed as previously described [5, 58]; this involves Sanger sequencing using a variety of different primers. Sequences were assembled from the resulting contigs and trimmed to the VP1 region using SeqMan Pro (Lasergene v.8 package, DNAstar Inc.). GenBank records were examined to exclude duplicates and isolates for which the year of sampling or country of origin were unavailable. Where two or more records from the same isolate were available, the more recently sequenced version was used. Sequences were aligned using the MUSCLE [42] plugin in Geneious 5.6.4 (Biomatters, Ltd.), and restricted to VP1 only.

2.2.2 Molecular clock and skyride analysis

BEAST [39] was used to conduct a molecular clock analysis using a GTR+I+G substitution model, a relaxed uncorrelated lognormal molecular clock [37], and a GMRF Bayesian skyride tree prior [102]. Multiple Monte Carlo Markov Chain (MCMC) runs of 100,000,000 states each and a burn-in of 10% were combined to obtain a set of 9,000 samples with effective sample sizes of at least 200 for all numerical model parameters. Tracer 1.5 (`http://beast.bio.ed.ac.uk/Tracer`) was used to reconstruct the skyride plot and investigate parameter estimates.

2.2.3 Phylogeography

A first phylogeographical analysis was performed using the posterior sample of trees from section 2.2.2 as an empirical tree set. An asymmetric rate matrix was assumed. Traits were selected depending on the status of the disease in the country of sampling as follows: for samples from areas in sub-Saharan Africa where FMDV is endemic, the country was used. However, each epidemic in North Africa and the Middle East was treated as a separate trait even where (in the case of Libya) more than one epidemic had occurred in a single country. As a previous analysis of the Egyptian sequences from 2012 [2] determined that the isolate EGY/2/2012 (designated as strain Alx-12) was most likely the result of a separate introduction to the other sequences from this outbreak (strain Ghb-12), these two lineages were also treated as different traits. This allowed investigation of the source of each epidemic, and the two Egyptian lineages, separately. As any given outbreak could not be the origin of an earlier one, the rates of transition between such states in this direction (e.g., from Egypt in 2012 to Libya in 2003) where set *a priori* to be zero. TreeAnnotator 1.7 was used to produce the MCC tree, with branches coloured by trait from this analysis.

Geographical movements were reconstructed using the Markov Jumps procedure [103] to give times of state changes along each branch of each tree in the posterior output. These were used to estimate a probability distribution for the country of origin of each of the epidemics, as follows: for every tree in the posterior sample, the tips corresponding to all the samples from an epidemic were identified and the node corresponding to their most recent common ancestor found (this was the tip itself in situations where only a single sequence was available for a given epidemic). If the reconstructed location state of this node was not the same as that of the tips, the epidemic was recorded as being the result of multiple introductions in this particular posterior sample. Otherwise, the reconstructed state change that took the lineage into the epidemic state was found, and the trait that was the origin of this jump was recorded. Summarising this information over all trees from the sample gave the posterior probability distribution of origins.

A second phylogeography analysis was conducted by restricting the dataset to only the 215 sequences from sub-Saharan Africa, in order to identify patterns of movement within the continent. For this purpose a second set of phylogenetic trees was produced, using the same molecular clock and tree prior as above. A separate phylogeographic analysis was performed, using this as an empirical tree set, with the Bayesian stochastic search variable selection (BSSVS) procedure used to identify pairs of countries for which the hypothesis that the rate of movement between them was nonzero was supported by a Bayes Factor (BF) value greater than 3. For this analysis, a symmetric rate matrix was assumed. QuantumGIS 1.8.0 (`http://www.qgis.osgeo.org`) was used to visualise well-supported nonzero rates on a map.

2.2.4 Host species analysis

A final discrete-traits analysis was performed to investigate transitions between different host species for the virus; the dataset was further restricted to those sequences from sub-Saharan Africa with an identified host. Information from GenBank records and the Picornavirus Home Page (`http://www.picornaviridae.com/`) was used to provide this information. The posterior set of trees from this was used for the host species analysis. Reconstruction of state changes was again performed using Markov Jumps, and the number of transitions between each pair of species was counted for all samples from the MCMC and summarised to give the median number of each type of host-to-host transmission taking place over the phylogeny, and the posterior probability that at least one event of each type occurred.

2.3 Results

2.3.1 The data

There were a total of 201 records for distinct isolates available in the NCBI Nucleotide database. (Information on the origins of the 49 newly-sequenced isolates can be found in table A.2, appendix A). The total dataset consisted of 250 sequences. All were 648 base pairs (bp) in length with the exception of five West African examples (all of topotype VI) which were each of 651 bp. Table 2.1 summarises the locations and years of sampling, and table A.1 describes the data in more. Since all relevant sequences were sampled prior to the partition of Sudan in 2011, the country was treated as a single location state for this analysis. 215 sequences were from sub-Saharan countries and the remaining 35 from outbreaks in North Africa and the Middle East.

Country	No. isolates	Date range
Angola	1	1974
Bahrain	5	2012
Botswana	6	1977-1998
Burundi	2	1986-1991
Cameroon	3	2000-2005
Côte d'Ivoire	1	1990
Democratic Republic of the Congo (or Zaire)	2	1974-1982
Egypt	22	2012
Eritrea	3	1998
Ethiopia	25	1990-2010
The Gambia	2	1979
Ghana	2	1990-1991
Kenya	65	1957-2007
Libya	5	2003-2012
Malawi	1	1975
Mozambique	3	1970-1983
Namibia (or South West Africa)	4	1989-1998
Niger	1	2005
Nigeria	2	1975-2007
North Yemen	1	1990
Palestinian Autonomous Territories	1	2012
Rwanda	4	1996-2004
Saudi Arabia	1	2000
Senegal	5	1975-2009
South Africa	31	1959-2001
Sudan (and South Sudan)[a]	6	1977-2010
Tanzania	2	1975-1986
Togo	1	1990
Uganda	13	1975-2007
Zambia (or Northern Rhodesia)[b]	6	1948-1996
Zimbabwe (or Rhodesia)	24	1972-2003
All sequences	**250**	**1948-2012**

[a]All isolates sampled before partition of country in 2011

[b]Isolate RHO/1/48, whose name suggests an origin in modern-day Zimbabwe, was in fact sampled in Northern Rhodesia, which is modern-day Zambia (see `http://www.picornaviridae.com/aphthovirus/fmdv/fmd_history.htm`)

Table 2.1: Countries and dates of sampling for available FMDV serotype SAT 2 isolates

2.3.2 Molecular clock and skyride analysis

Figure 2.2 is the maximum clade credibility (MCC) tree from this analysis, with branches coloured by topotype. The year of the median time of most recent common ancestor (TMRCA) for all sequences was 1879, with a 95% highest posterior density (HPD) interval from 1849 to 1905.

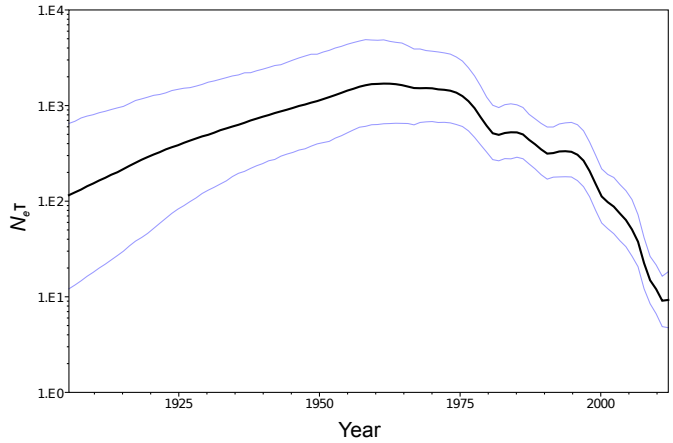


Figure 2.1: GMRF Bayesian skyride plot of $N_e\tau$ (effective population size times generation time) against calendar time. Blue lines are the boundaries of the 95% highest posterior density interval.

The estimated parameters (posterior medians) of the molecular clock were a mean of 2.5×10^{-3} substitutions per site per year (95% HPD: $1.85 \times 10^{-3} - 3.29 \times 10^{-3}$) and a standard deviation of 2.95×10^{-3} ($1.55 \times 10^{-3} - 4.64 \times 10^{-3}$). (Normal BEAST output, counter-intuitively, gives the mean of the lognormal distribution of clock rates on the real scale and the standard deviation on the log scale, and these are the numbers that are generally reported in papers. I depart from this and give both on the real scale, here and in subsequent chapters.) The reconstructed skyride plot can be seen in figure 2.1. Genetic diversity peaked around 1965 and then began to decline, at a rate becoming faster around 1995.

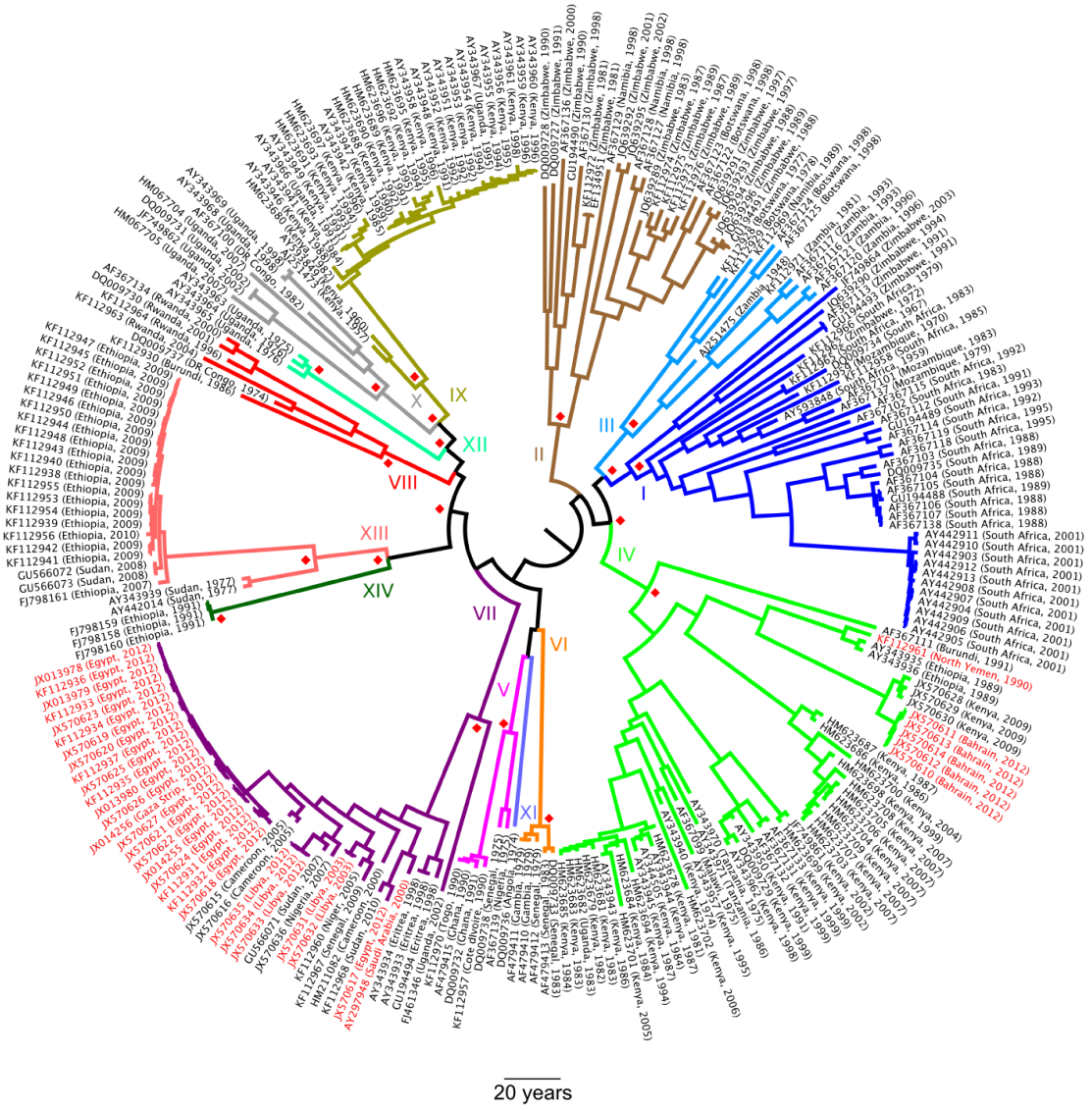


Figure 2.2: Maximum clade credibility tree of all sequences included in the dataset. GenBank accession numbers and countries and dates of sampling are given at the tips; sequences isolated during epidemics in North Africa and the Middle East are in red. Branches are coloured and labelled by topotype (I-XIV). Red diamonds indicate clades with posterior probability >0.9 (within topotypes, these are omitted for all nodes except for the common ancestor of the topotype).

2.3.3 Phylogeography

Figure 2.3 displays the MCC tree, with branches coloured by location of sampling for tips and highest posterior probability location for internal nodes. For clarity, sub-Saharan countries have been grouped by UN region.

Figure 2.4 gives the posterior distributions for the country that seeded each North African and Middle Eastern epidemic occurring since 2000. No epidemic other than the one in Egypt and Palestine in 2012 was reconstructed as the result of multiple introductions in any sampled MCMC state. Kenya was overwhelmingly the most likely origin for the 2012 Bahrain epidemic (posterior probability 0.89), as was Cameroon for the Ghb-12 lineage of the 2012 Egypt/Palestine outbreak (posterior probability 0.81). Results were less decisive for the other five outbreaks, with no origin having a posterior probability of more than 0.6. In particular, while the Egyptian Alx-12 lineage appeared most likely to be a descendant of a Sudanese isolate (posterior probabillity 0.6) it was also closely related to the virus from 2000 in Saudi Arabia (posterior probability 0.21). Also notably, there was practically no suggestion that any of the 2012 outbreaks were the direct descendants of each other.

The map in figure 2.5 displays BF>3 supported nonzero rates of transition between endemic countries, from the separate analysis of sequences from these only. Most identified links were across a shared land border; longer-distance links were usually in cases where there were intervening countries from which samples were not available. Longer links also tended to have lower BF support.

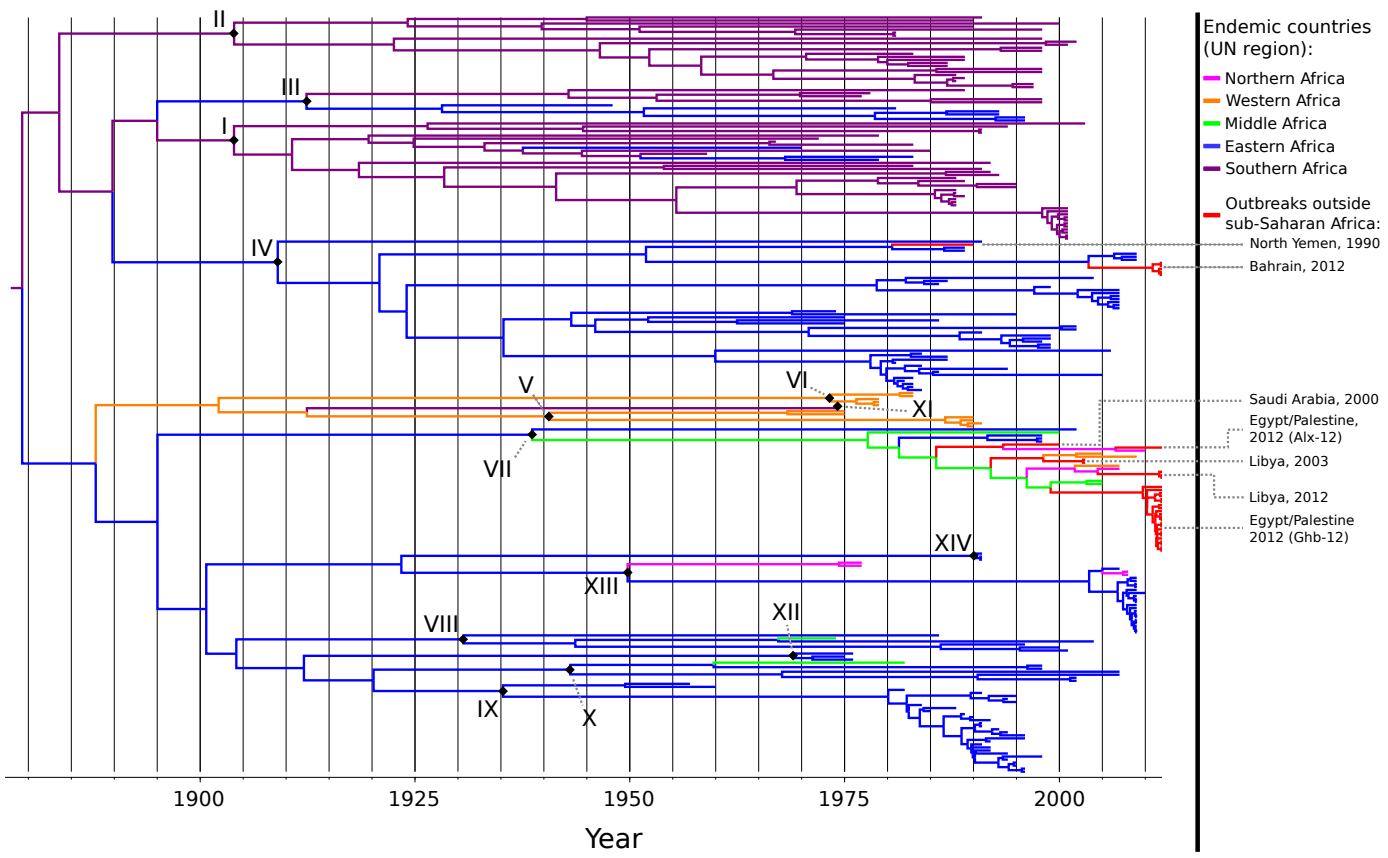


Figure 2.3: Maximum clade credibility tree of all sequences; branches are coloured by UN region within sub-Saharan Africa or red for outbreaks in other areas. Roman numerals and black diamonds indicate nodes representing the common ancestor of each topotype, or, where only one sequence was available for that topotype, the tip corresponding to that sequence.

		Destination		
		S. caffer	**Cattle**	***A. melampus***
Origin	***S. caffer***	-	10 (1.00)	3 (0.85)
	Cattle	5 (0.94)	-	0 (0.48)
	A. melampus	6 (0.88)	1 (0.53)	-

Table 2.2: Median (across all trees in the posterior sample) numbers of reconstructed Markov Jumps between each pair of species in the host species analysis. (This posterior sample of trees is summarised by figure 2.6.) Numbers in brackets are posterior probabilities for at least one such jump having occurred since the time of common ancestor of the 168 isolates.

2.3.4 Host species analysis

Only 169 sequences had an identified host, which was *S. caffer* in 28 cases, domestic cattle in 130, *A. melampus* in 10, and a pig in 1. The latter was excluded as a single example was unlikely to be adequate for the purpose of investigating the sources of infections in pigs. Figure 2.6 shows the MCC tree. Branches are coloured by host; clades representing topotypes are annotated with a diamond. The most likely root state (the host species of the common ancestor of all known SAT 2 isolates) was *S. caffer* with a posterior probability of 0.53.

Table 2.2 summarises the results of a Markov Jumps analysis for changes of host species. The median number of jumps across all trees in the posterior are given for each pair of hosts, along with the posterior probability that the total number of such transitions was nonzero. The median number was nonzero in all cases except transitions from cattle to *A. melampus*, but the only type of transition for which there was 95% support for at least one such jump occurring was from *S.caffer* to cattle.

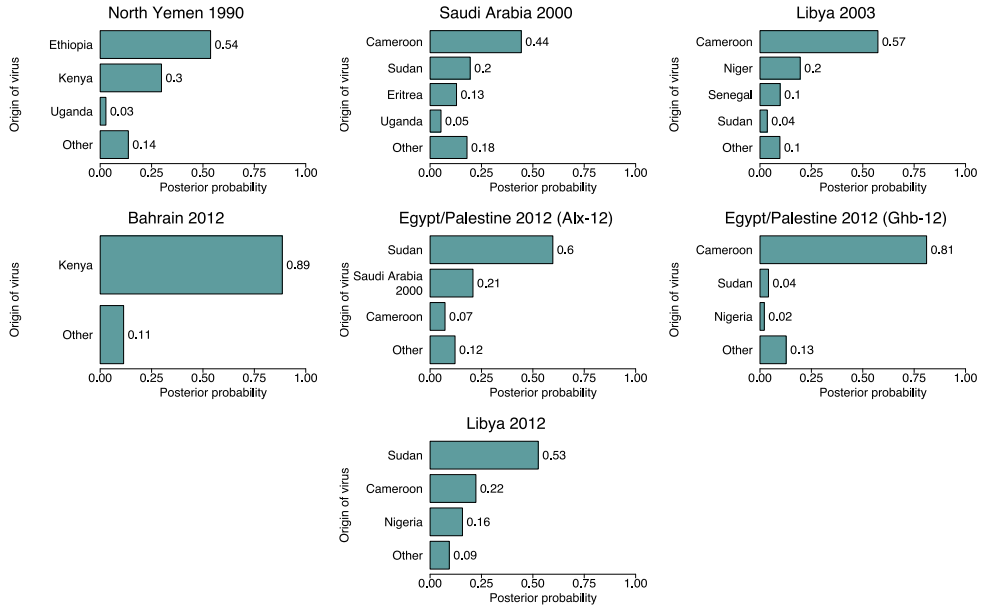


Figure 2.4: Posterior probability distributions for the countries or epidemic states that were the origins of reconstructed Markov jumps seeding SAT 2 outbreaks in North Africa and the Middle East, 2000-2012. Only origins with a posterior probability of 0.02 or more are shown individually.

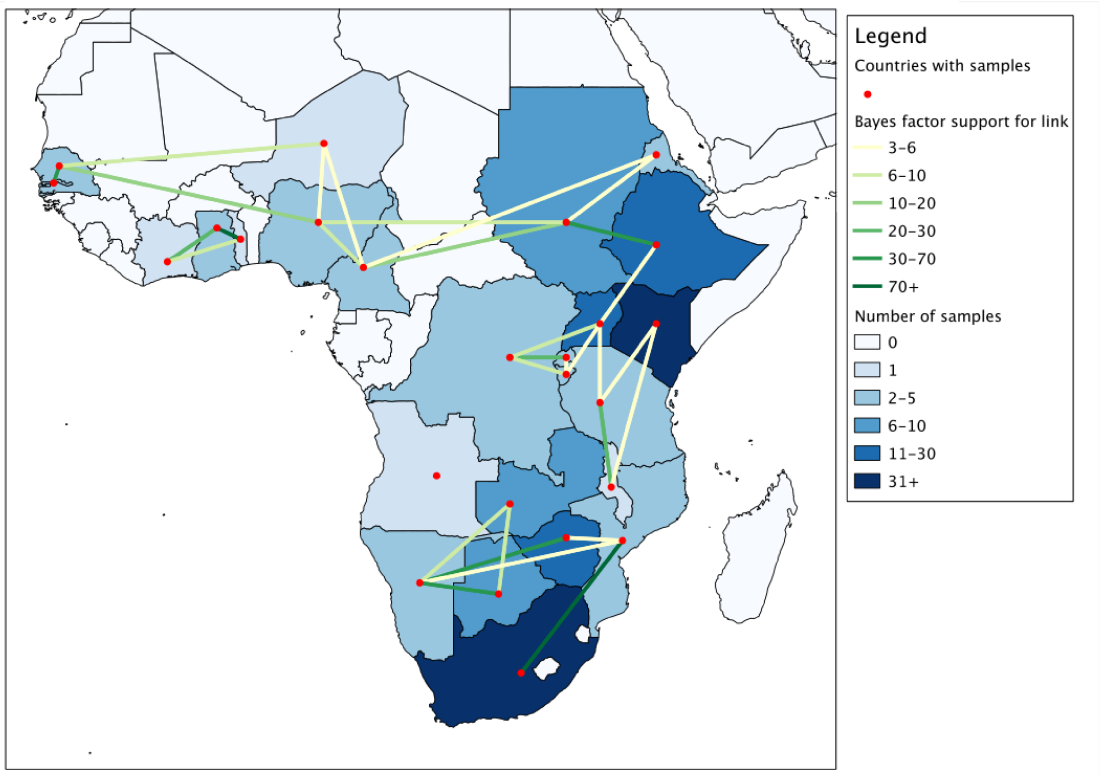


Figure 2.5: Map of Africa demonstrating links between countries with Bayes factor (BF) support >3 identified from the BSSVS analysis. Countries are coloured by number of sequences available from that location; links are coloured by BF value.

2.3.5 Accession numbers

The 49 newly determined sequences are now available in GenBank with accession numbers KF112928 to KF112976.

2.4 Discussion

In this chapter, I applied some of the methods of phylodynamics and phylogeography to the VP1 gene sequences of all SAT 2 isolates available at the time of

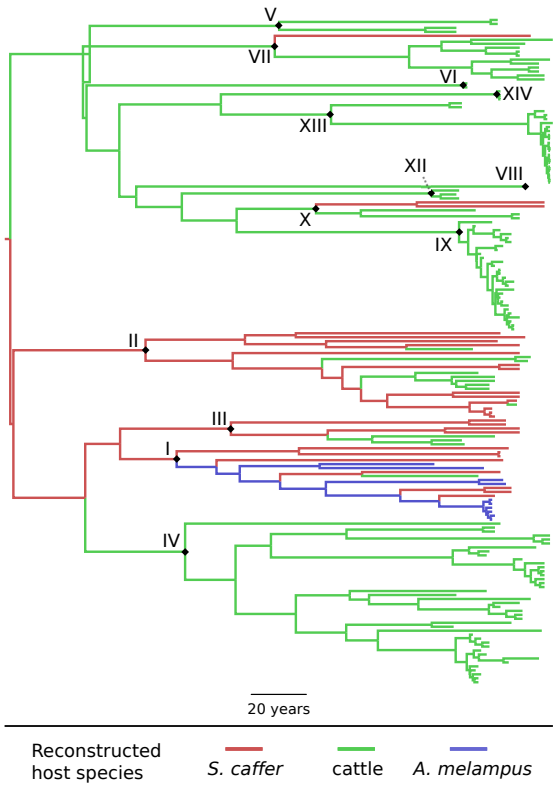


Figure 2.6: Maximum clade credibility tree of 168 sequences coloured by reconstructed host species. Branches are coloured by host. Roman numerals and black diamonds indicate nodes representing the common ancestor of each topotype or, where only one sequence was available for that topotype, the tip corresponding to that sequence.

writing. It has some limitations, largely imposed by the nature of the available data. The sampling is effectively opportunistic and markedly unbalanced, and the exact effect of this on the skyline and discrete trait inference methods used here is a topic which I investigate in its own right in chapter 3. This makes the results of the host species analysis in particular somewhat incomplete, firstly because very few countries have available sequences from both cattle and wild animals, and secondly because no sequences at all are available from sheep or goats, despite the hypothesis that they play an important role in the maintenance of FMDV populations [20]. In addition, use of just the country of origin as a location state

gives coarse resolution; lack of links between locations may be simply the result of lack of sampling in areas sufficiently close to the relevant borders, but restricting to only those sequences for which more detailed location information is available would have greatly decreased the size of the dataset.

The VP1 segment was used simply because it has been the most commonly sequenced section of the genome, but use of a larger part would be more suitable and is now more viable in the era of next-generation sequencing. At the time of writing there are nine publicly available sequences for the full SAT 2 genome, and an additional seven for the full coding region (polyprotein gene). While recombination within the structural protein region (VP1-3) appears to be rare, and thus should not be a cause of concern in interpreting this analysis, it is widespread in other parts of the genome [22, 72]. This likely renders a naive whole genome phylogenetic approach inadvisable. Indeed, van Rensburg and Nel [150] found that the leader and 3C proteinases of SAT FMDVs displayed very different branching patterns to the VP1, and it is this recombination that likely explains the findings by Yoon et al. [170] and Lewis-Rogers et al. [97] that, when a full-genome analysis is performed, the SAT strains do not form separate clades. However, while the entire genome may not be a good subject for analysis, future work could use the whole structural protein region, rather than just VP1.

The estimated substitution rate of 2.5×10^{-3} substitutions per site per year is very similar to the 2.48×10^{-3} given by Tully and Fares [145] for their analysis on the VP1 segment of all FMDV serotypes, but considerably faster than their specific estimate for SAT 2 of 1.07×10^{-3}, and the 95% HPD interval of 4.90×10^{-6} to 1.14×10^{-3} given there does not overlap with the one found here. That estimate, however, is very imprecise compared to the results for all other serotypes in the same paper, and the dataset of 32 sequences used by the authors was also much smaller than mine, covering only ten of the fourteen topotypes. The slower substitution rate estimate in that paper naturally corresponded to an earlier

estimated TMRCA of the year 1777, with a 95% HPD interval from 1747 to 1913, also very different to the estimate here, although in this case the HPD intervals do overlap. Yoon et al. [170] also estimated a slower overall substitution rate (1.46×10^{-3}) for all serotypes, but this analysis was on the full genome, ignoring recombination, and a different rate might be expected. That paper also estimated a much earlier TMRCA for SAT 2 in 1615, with the 95% HPD interval from 1324 to 1866, slightly overlapping mine.

The decline in genetic diversity of FMDV in the latter part of the twentieth century has previously been noted by Yoon et al. [170], whose analysis of all seven serotypes also identifies a peak in the middle of the century and a faster decline starting around 2000. A similar peak was also identified by Tully and Fares [145], although their analysis suggests a subsequent sharp *increase* in the last years of the century. A potential explanation for the mid-century decline is the vaccination and fencing measures that have been put in place over the past decades in southern Africa in order to prevent the infection of cattle by wild animals [71, 156]. The steeper decline observed starting in the mid-90s may be a sampling artefact due to the inclusion of a disproportionate number of sequences from comparatively well-sampled epidemics with dates from this time period, as the increased number of coalescent events associated with such data tends to lead to the reconstruction of spurious population bottlenecks (see chapter 3). Alternatively, it could reflect a genuine decrease in diversity, possibly due to improving farming practices.

As FMDV in Africa is presumably generally spread overland by animal movements, the inference of a particular country as the origin of a particular epidemic in this analysis should not be interpreted as the last country that the lineage was present in before the start of the epidemic; for example no strain could have moved directly in this way from Cameroon to Egypt or Libya for the obvious reason that there are intervening countries on any route between them. Instead, this analysis provides a probability distribution of the location of the most recently

observed ancestral lineage to that which gave rise to the outbreak; no conclusions can be drawn regarding the route that might have been taken to get from one to the other. In particular, the wide distribution of topotype VII, from Nigeria to Eritrea, has previously been noted by Bronsvoort et al. [19] and is thought to be the result of extremely long distance cattle movements which are known to occur between Cameroon and Sudan. Thus, although the origins for the Libya 2003 and Ghb-12 outbreaks are suggested to be Cameroon, they could well have first made their way east to Sudan before crossing the Sahara, with Sudan not identified as their origin because strains more closely related to them than known Sudanese isolates have never been sampled in that country.

Three separate SAT 2 outbreaks in North Africa and the Middle East in a single year, nine years after the last such recorded event, might seem unlikely to be independent events, but the evidence here adds further weight to the suggestion [2] that these were not the result of a single introduction and that the concurrence is due to coincidence or regional circumstances that have made such events more likely. If the latter, this situation may not be particular to SAT 2: a new serotype A virus with a probable origin in sub-Saharan Africa was also discovered in Egypt in 2012 (`http://www.wrlfmd.org/fmd_genotyping/2012/WRLFMD-2012-00011\%20A\%20Egypt\%202010-2012.pdf`), although whether this was a genuinely new introduction in the very recent past or its detection was the result of heightened surveillance as a result of the ongoing SAT 2 emergency seems an open question, given the fairly frequent occurrence of serotype A in the country [83, 86].

The two topotype IV outbreaks, North Yemen 1990 and Bahrain 2012, were determined to have Kenya or (in the former case) Ethiopia as likely origins. The Bahraini isolates came from cattle that had been recently imported from Saudi Arabia (`http://www.promedmail.org/direct.php?id=20120507.1125683`). It is unlikely that these strains arrived in the Middle East directly from Kenya by sea; Di Nardo et al. [33] describe cattle movement patterns in the region and did

not identify such exports. They do, however, identify imports to Yemen and Saudi Arabia from Somalia, a country whose SAT 2 strains have never been sequenced. Type O FMDV outbreaks in Yemen have previously been traced to cattle from eastern Kenya and Ethiopia traded through markets in Somalia [33], so this would seem the most obvious explanation. Identification of which SAT 2 topotypes are in fact present in Somalia would help confirm this. If the 1990 outbreak originated in Ethiopia then another possible export route would go through Djibouti.

The Alx-12 strain identified in Egypt is genetically distinct from Ghb-12 and the Markov Jumps reconstruction suggests that the most likely origin country was Sudan, but that it could also be descended from the 2000 Saudi outbreak. As it is highly unlikely that both Alx-12 and Ghb-12 were the product of a single viral lineage arriving in Egypt, it seems most probable that there was indeed a fourth 2012 viral escape of this serotype from sub-Saharan Africa. While I did identify different most likely countries of origin for the two strains, this does not rule out the introductions being the result of the import of the same group of infected animals from Sudan, as the Ghb-12 lineage, originating in Cameroon, may have travelled east on its route to Egypt. If there were indeed two separate introduction events, the cause might be the increase in cattle imports to Egypt identified by Kandeil et al. [78].

The close relationship of Alx-12 to the Saudi strain does, however, suggest another possibility: that this lineage may have been present but undetected in North Africa and the Middle East since 2000 or even earlier, its detection in 2012 being the result of the increased surveillance connected to the Ghb-12 outbreak. Since other FMDV serotypes are endemic in these areas [87, 126], it is plausible that it was overlooked. In this scenario the virus persisted in the region following the 2000 outbreak, or even was present before that. If true, then the virus is likely to have been maintained in sheep or goats, species in which clinical disease is less likely to be apparent [20]. Sheep populations have previously been implicated

in maintaining FMDV in these areas [122, 126]. Further viral samples from the area and from other countries where topotype VII is present would be required to clarify the picture. A question that also arises is why, in this case, the Ghb-12 introduction would cause a rapidly-spreading epidemic and disease control emergency while the existing presence of Alx-12 did not.

Aside from the clear difference between Alx-12 and Ghb-12, there was no suggestion that any other outbreak was the result of multiple introductions, and none of the 2012 outbreaks was suggested to be the source of any of the others.

It is generally accepted that FMDV is spread locally in Africa by movements of both livestock and wild animals (that it is frequently subclinical in wild *S. caffer* is considered a major challenge to control of the disease [4, 8, 144, 157]). The phylogeographical analysis within endemic countries presented here lends some formal support to this hypothesis, as movements over large distances were rarely indicated except where there were intervening countries from which no samples were available, and where such links were suggested the BF support was usually on the low side (as in the links from Malawi to Kenya, and Mozambique to Namibia). Investigation into whether the long-distance links between Cameroon and Nigeria and more distant countries to both the west and the east are genuine would require sequences from intervening nations, which are currently unavailable for the full VP1 gene. However, as mentioned above, the close relationship between sequences from Cameroon and samples from Eritrea and the 2000 Saudi outbreak was previously noted by Bronsvoort et al. [19], who point out that cattle are indeed traded directly from Sudan to Cameroon and could have carried the virus over this distance. At the time no sequences from Sudan or the Central African Republic were available, so the authors acknowledged they were unable to conclusively demonstrate this. The picture remains patchy, but this analysis does include Sudanese sequences and links from Cameroon to both Eritrea and Sudan are supported, providing some further evidence for this hypothesis.

Because of the geographical distribution of the available sequences, much more information is available for countries in eastern or southern Africa than for western and central areas, where the picture is fragmentary at best. The situation in the countries south of Cameroon is particularly unclear; apart from sequences from the DRC that are most closely related to isolates from its east, the only isolate from this region is a single Angolan example from 1974, the unique available sequence from topotype XI. No strains from Equatorial Guinea, Gabon or Congo have ever been sequenced. Whether topotype XI still exists, and more generally what the status of SAT 2 is in this region, would appear to warrant further investigation.

The situation in West Africa is better; there are in fact around fifty sequences from countries from Cameroon westwards for partial sections of the VP1 gene that were ineligible for this analysis due to being insufficiently long. Sangaré et al. [127] performed an initial phylogenetic analysis on most of these; an extension to this analysis could apply the same phylogeographical methods to the shorter sequences from this area only.

As mentioned above, the host species analysis should be interpreted with caution due to the incomplete nature of the sampling. While there is not strong support here for the hypothesis that virus escapes from natural parks in southern Africa are the result of impala jumping fences [144, 157], the only available *A. melampus* sequences are from the Kruger National Park in South Africa and few subsequent cattle sequences are from any country adjacent to the park. While the colouring of branches in figure 2.6 indicates the most probable host for the common ancestor of each topotype, this is unlikely to be reliable as many topotypes have only had isolates sequenced from cattle, yet there is no reason to believe that they do not also infect buffalo. The role of any other hosts, such as sheep, cannot be investigated. Nevertheless, that SAT 2 originated in *S. caffer* is consistent with the consensus that buffalo are the maintenance host for the SAT strains [8]. Subsequent transitions from *S. caffer* to cattle are reconstructed with support

at the 95% level for the count being nonzero; this is consistent with previous literature implicating buffalo as the cause of epidemics in southern Africa [158]. Transitions from *S. caffer* to *A. melampus* and vice versa, and cattle to *S. caffer* are also frequently reconstructed with considerable posterior support for their occurrence, but not reaching the 95% level. That transitions from buffalo to impala, at least, must occur is generally accepted [8, 157]. It is also feasible that cattle and impala infect buffalo, but that hypothesis is not necessary to explain the epidemiology of the virus.

In summary, this chapter used up-to-date methods and sequence data to update the picture of the behaviour of the SAT 2 serotype on a continental level. Support is given for generally accepted characteristics the virus: that it is spread over generally short distances by the land movements of infected hosts, and that African buffalo are an important maintenance host. The previous consensus that the 2012 outbreak strains are unrelated and probably did not have the same origins has been strengthened by a formal phylogeographical analysis. Evidence is also provided that the decline in FMDV genetic diversity in the latter part of the twentieth century applies to this serotype. Future work on this virus would be enabled by further sequencing, perhaps of a larger part of the genome, with a more methodological sampling scheme. It is the question of the latter which I now turn to.

Chapter 3

The effects of sampling strategy on the quality of the reconstruction of temporal and spatial dynamics using genetic data: a simulation study

3.1 Introduction

The quantity of available genetic data on pathogens is already very large, and will only grow in future. The days when, in performing a phylogenetic analysis, it might be appropriate to use every sequence available simply as the result of scarcity of data are long gone for some pathogens and cannot last long for many others. This raises the important question of how, in future, a set of sequences should be selected for analysis. As I outlined in chapter 1, there are two concerns here. Firstly, only very basic or approximate phylogenetic methods can analyse thousands of sequences in reasonable computational time. Subsampling is in

many cases a necessity. Secondly, with large amounts of genetic data now readily available, the biasing effects of a particular choice of sample must be considered. This has long been an important consideration in epidemiological studies; however, in molecular epidemiology it has lagged as a concern, probably because in the past, genetic data of any sort has been at a premium. This must now start to be remedied.

In this chapter, I have performed a large-scale simulation exercise to determine the effect of different sampling schemes on the reconstruction of spatial and temporal dynamics of pathogen populations. For the reconstruction of temporal dynamics, this involved the use of the GMRF Skygrid plot [52]. This is the most recent iteration in the Bayesian skyline family of methods [36, 102], which use coalescent methods to infer past variation in a product $N_e\tau$, which is the product of the effective population size (EPS) N_e and the time between generations τ. Usually, no specific generation time is assumed and instead the product is estimated. For brevity, when this chapter refers to "EPS" it actually refers to this product. Unlike simple, parametric models of EPS (common examples of which are constant size, exponential growth, or logistic growth), the members of the skyline family are non-parametric: the timeline is divided up into a finite number of intervals, and the EPS is assumed to be constant on each interval but can change between them. Each value of the EPS on each interval is estimated along with the phylogeny.

While coalescent-based methods were originally conceived with populations of organisms in mind, such that the EPS is the (effective) number of individuals and the generation time the time between births, in studies of pathogens it has often been interpreted in an epidemiological sense, so that the population is of infected individuals and the generation time the serial interval. This has been shown to be mathematically inaccurate [48, 153]; coalescence rates under an epidemiological model are governed by both incidence and prevalence and cannot generally be used to infer prevalence alone. For this reason I prefer to regard populations of interest

as being made up of pathogens. Skyline inference also makes the assumption that lineages form a single, freely-mixing population. For this reason, and also because the effective and census population sizes will rarely be the same, the numerical values of the estimates of the EPS do not literally refer to a number of individuals, and exact interpretation of them is generally not attempted. Instead, temporal trends are generally examined for evidence of changes in population dynamics over time.

The assumption that lineages are part of a single, freely-mixing population will always be violated in practice. Structured coalescent models, which subdivide this population into freely-mixing "demes" and allow lineages to transfer between them, are well-developed [106]. However, current implementations of these in phylogenetics packages assume that the size of each deme is constant over time [30, 152]. As historical changes in population size are of considerable epidemiological interest, the skyline family is often still used, despite the fact that a key model assumption is generally invalid. One aim of this chapter is to investigate the effects of this discrepancy.

For the reconstruction of spatial dynamics, I used a discrete traits model [94], which is commonly used to investigate the spread of infectious organisms between geographical locations (as in chapter 2). Location is treated as an extra position in the genetic alignment, evolving according to the same continuous-time Markov chain (CTMC) process that is used to model mutation. Lineages are in one discrete location or another, and transition to others occurs according to a rate matrix whose entries are estimated along with the phylogeny. Using the Bayesian stochastic search variable selection (BSSVS) procedure, hypothesis tests can be performed to identify rates whose entries are nonzero, and those are often plotted on a map to display the geographical spread of the infection.

The effects of sampling strategy on phylodynamic inference is a neglected area

of study, which has been identified as an important future problem [47]. Only two previous papers have investigated the accuracy of temporal reconstruction in the context of infectious disease. Both simulated epidemics according to a mathematical model of transmission, and subsequently used the coalescent-based skyline model to reconstruct the dynamics. Stack et al. [137] applied this to a cyclical epidemic, intended to be analogous to measles, and found that the accuracy of the reconstruction was greatly improved if sequences were sampled either during the decline in case numbers that follows an epidemic peak, or in the trough of cases immediately following it. The work of de Silva et al. [31] concentrated on strategies for analysing the early stage of an epidemic, while it was undergoing exponential growth. They found that the reconstructed dynamics almost always indicated a wholly spurious flattening off of the epidemic in the period prior to the date of the last-collected sample, and that this effect was worse, with the tailing off occurring earlier, if one sample was taken per generation of infection rather than if the number of samples taken in a generation was proportional to the logarithm of the number of available samples. They also showed a flattening effect when samples taken were epidemiologically related to one another.

Both the aforementioned papers reconstructed dynamics from simulations in which the free-mixing assumption of the skyline family was not violated. To my knowledge, no published simulation study on the effects of sampling schemes has explored the effect of population structure in an infectious disease context. However, there are examples from the literature of eukaryotic phylogenetics. The most important difference between a study of that sort and an analysis appropriate to a pathogen study is that in the former case, the period between the collection of samples is regarded as negligible compared to the evolutionary timescale, and as a result all tips of the tree are treated as contemporaneous. This makes any consideration of the temporal nature of the sampling scheme irrelevant. Chikhi et al. [23], using a simple two-step coalescent model rather than a member of the

skyline family, noted that spurious population bottlenecks tended to be detected if the sampling scheme was such that some subpopulations (demes) were missing. Heller et al. [64] employed the skyline in a similar context and found similar results; a spurious population bottleneck immediately prior to the time of sampling was commonly found if samples were not taken from every deme. Notably, both papers assumed a constant total EPS. An obvious question is what the effect of temporal sampling schemes on structured populations is, and what happens when EPSs are allowed to vary.

Research on the impact of sampling on the reconstruction of spatial dynamics is even scarcer. The only example I am aware of is a very recent paper by De Maio et al. [30], which showed that rates of transition between locations tend to be underestimated due to a conceptual problem with the CTMC model, which I address in more detail under Discussion.

The earlier studies on temporal reconstruction simulated phylogenies and sequences using an epidemic model [31, 137], comparing parameter values from this to EPS estimates from a coalescent model. Volz [153] demonstrated how to simulate phylogenies under a coalescent process whose underlying dynamics were a potentially complex model of transmission. Nevertheless, I chose to simulate under a coalescent process in a population whose EPS obeyed a given function directly. I did this for three reasons. Firstly, because this is the model under which the skyline-family models perform their reconstructions. Secondly, because in the Volz solution, prevalence through time is derived as a function of birth and movement rates and I could not easily pick any function of interest to represent the "true" dynamics. Finally, because the primary focus of this exercise was to investigate the effect of sampling scheme on the investigation of the global dynamics of an endemic disease (such as FMDV in many parts of the world). Constructing an epidemiological model for the behaviour of an endemic pathogen on a global scale raises many questions that are out of the scope of this chapter; I

found it preferable to use an established model for the population dynamics of a collection of organisms. So, while the exact relationship between the reconstructed EPSs from a coalescent model and the dynamics of infection are complex [48, 153], I assume that such a relationship can in fact be quantified, and deal only with the effective size of the pathogen population. The demographic functions here were thus not intended to follow any particular model of disease dynamics; I instead investigated the quality of the reconstruction for various scenarios of variation in population size.

The finding of Heller et al. [64] that sampling from some populations and not others can falsely suggest population declines in reconstructed dynamics is pertinent to infectious disease studies, as it is a quite common practice in molecular epidemiological studies to analyse a large number of sequences recently collected as part of a single study with a more sparsely sampled dataset from other locations and times for comparison. The work of chapter 2, indeed, was an example of this; an overly large proportion of the samples were taken from the 2011 Egyptian FMDV epidemic, and I did indeed see a decline in the reconstructed EPS in the period immediately before the last samples were acquired. This pattern can be seen in other studies of, for example, influenza A virus [98], West Nile virus [110], and peste des petits ruminants virus [109]. It makes intuitive sense that the population structure might confound the analysis in this case; under the assumption of random mixing, if a large number of lineages coalesce very rapidly before sampling, it would suggest a small total population size, but if the population was in fact structured (as will always be the case in reality) and these samples were all taken from the same place this would be misleading as they would coalesce only with those from the same deme. Nevertheless, this has not been explicitly demonstrated in a population analogous to a population of pathogens with non-contemporaneous temporal sampling.

3.2 Methods

3.2.1 Sequence simulation

Datasets of 50,000 sequences were simulated under eight separate demographic scenarios. This was done in two steps: firstly, an overall "master" phylogeny was simulated under a coalescent process, and secondly the master phylogeny was used to generate sequences by simulating mutation along its branches.

The first four scenarios modelled coalescence occurring in an unstructured population of freely-mixing haploid individuals. The EPS, $N_e\tau$, in each population varied with a deterministic function $N(t)$. All phylogenies were simulated using custom Java code, making use of the existing classes for handling trees that are implemented in BEAST [39]. The master phylogeny for 50,000 simulated isolates was constructed by, firstly, randomly placing 50,000 tree tips over a period of 10 time units; the units t were intended to represent years and will be referred to as such hereafter. The 10 years were divided into 10,000 intervals and each sequence was assigned in turn to an interval with probability proportional to the function N evaluated at the midpoint of the interval. The exact sampling time was then selected by a draw from the uniform distribution with bounds confined to that interval.

With all tips placed, coalescence was simulated until one lineage remained. As some of my scenarios included functions N for which the distribution of coalescence times is not analytically tractable (sine waves, for example), I approximated each N by a step function N', with the value of N' on a step being equal to the value of N at the midpoint of that step. Formally, if the step size is a, $N'(t) = N(a \times \lfloor \frac{t}{a} \rfloor + \frac{a}{2})$. The value of a was set to 0.001 years throughout. The simulation therefore assumed a constant EPS during each step, and in that situation the time taken

for K lineages to experience one coalescence is exponentially distributed with rate $\frac{K(K-1)}{2N'(t)}$ [119]. I now switch to a backwards timescale s whose zero point is the time at which the final tip (or tips) were placed; suppose a function c converts between this timescale and the original forward one that N and N' were defined on. Suppose $L(s)$ is the number of extant lineages at time s; $L(0)$ is the number of tips with a sampling time of 0. The phylogeny was constructed according to the following algorithm, starting at $s = 0$ with no tips yet added to the tree:

1. If $L(s) = 1$, let s' be the earliest time of sampling amongst the tips that have not yet been added to the tree, and suppose there are m tips with this time of sampling. If no such tips remain, then stop; all lineages have coalesced. Otherwise, extend the single extant lineage from s to s'. Add the m tips to the tree and increase the lineage count; $L(s') = L(s) + m$. Then set s to s' and go back to the start.

2. If $L(s) > 1$, draw a coalescence time c from an exponential distribution with rate $\frac{L(s)(L(s)-1)}{2N'(c(s))}$.

3. Let $s_1 = c^{-1}(a \times \lfloor \frac{c(s)}{a} \rfloor)$; s_1 is the time of the first change point in the step function after s. Let s_2 be earliest time of sampling amongst the tips remaining to be added (if there are any). If $s + c < \min\{s_1, s_2\} = s'$ then continue all lineages extant at s to $s + c$ and then coalesce a random two of them; $L(s + c) = L(s) - 1$. Update s to $s + c$ and go back to the start.

4. We must now have $s + c \geq s'$. Continue all lineages extant at s on to s'. If $s' = s_2$, i.e. the time of the tip or tips after s is before the time of the end of the step, and there are m such tips, add them to the tree, so $L(s') = L(s) + m$. Then update s to s', and go back to the start.

The scenarios in which an unstructured population was used were as follows:

- **Scenario 1:** A population of constant size: $N(t) = 10$.
- **Scenario 2:** A population undergoing exponential growth in size: $N(t) = e^{0.361t}$. (The value of the exponent was chosen so that the integral of the function over the 10 time units is approximately 100, in common with other scenarios.)
- **Scenario 3:** A population whose size underwent shallow oscillations: $N(t) = 10 + 5(\sin\ t)$.
- **Scenario 4:** A population whose size underwent oscillations of greater amplitude and frequency: $N(t) = 10 + 7.5(\sin\ \pi t)$.

The remaining four scenarios assumed a structured population, and trees were simulated under a structured coalescent. This involved a finite number of demes $D_1, \ldots, D_n$, and the EPS within each deme varied according to functions $N_1, \ldots, N_n$; these were approximated by step functions $N'_1, \ldots, N'_n$ as before. A set of rates M_{ij} determined movement between demes, such that M_{ij} is twice the rate per year at which a lineage in deme D_i will move to deme D_j [161]. For convenience say $M_{ii} = 0$ for all i. When simulating, tips were first assigned to a time interval as above, based on the total population size across all demes at the midpoint of the interval. They were then assigned to a deme with probability proportional to the EPS of that deme at that midpoint, and then to an exact time point within the interval as before. The modification to the algorithm to construct the tree requires that the number of extant lineages in D_i at time s, $L_i(s)$, be kept track of. In the unstructured population only a single stochastic event, the coalescence of two lineages, could occur. In a structured population there are potentially $n + n(n-1)$ types of event; n coalescences within a deme and $n(n-1)$ movements between demes. At a time s, following Wakeley [161], I temporarily converted to a timescale in which all times were scaled by the total EPS $N(s) = \sum_{i=1}^{n} N_i(s)$, and define $c_i(s) = \frac{N_i(s)}{N(s)}$, i.e. the fraction of the total

population contained in D_i. The rate in this timeline at which two of $L_i(s)$ lineages in D_i coalesce is $\frac{L_i(s)(L_i(s)-1)}{2c_i(s)}$ and the rate at which one of the $L_i(s)$ lineages moves to D_j is $\frac{L_i(s)M_{ij}}{2}$. The time to *any* stochastic event after s is therefore exponentially distributed with rate R where:

$$R = \sum_{i=1}^{n} \left[\frac{L_i(s)(L_i(s)-1)}{2c_i(s)} + \sum_{j=1}^{n} \frac{L_i(s)M_{ij}}{2} \right] \tag{3.1}$$

Step 2 of the algorithm above is modified to draw a time c from this distribution. I then scaled c back to the main backwards timeline by multiplication by $N(s)$. As before, if $cN(s)$ is after the introduction of new tips, or a change of EPS determined by the step function, I move to the time of that event and no coalescence or migration occurs. Otherwise, the exact event that happens is determined by choosing one with probability proportional to the corresponding component rate in (3.1). I move forwards to $n + cN(s)$ and then, if the event is a coalescence, coalesce two lineages in the appropriate deme, or if it is a migration, move a lineage from one deme to another.

The population structure used is depicted in figure 3.1. The circles represent six demes $D_1, \ldots, D_6$; two small (D_1 and D_4), two medium (D_2 and D_5) and two large (D_3 and D_6). The exact relative sizes of these varied depending on the scenario. Arrows represent nonzero M_{ij}. Movement rates are symmetrical and invariant over time in all scenarios; thick arrows represent a rate of 0.05 per lineage ($M_{ij} = 0.1$) in the source population per year between the respective demes; thin arrows 0.025 ($M_{ij} = 0.05$). In this way, there is movement between each deme and four of the five others, two at a fast rate and two at a slow one. Let $s_i = i \pmod 3$. The demographic scenarios considered were as follows:

- **Scenario 5:** A structured population of constant size. $N_i(t) = 10s_i/12$. Hence $\sum_{i=1}^{6} N_i(t) = 10$ and the total EPS is the same as in case 1.

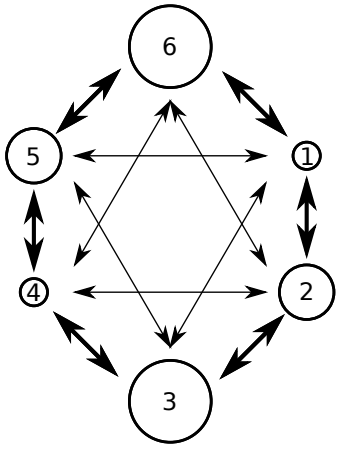


Figure 3.1: Depiction of the population structure used in structured coalescent simulations. Circles represent demes; two are small, two medium and two large. Thick arrows represent fast rates of movement between demes (0.05 transitions per lineage per year) and thin arrows slower rates (0.025 per lineage per year).

- **Scenario 6:** A structured population in which the size of each deme experiences gentle oscillations, and the oscillations are all in sync: $N_i(t) = s_i(10 + 5(\sin\ t))/12$. Hence $\sum_{i=1}^{6} p_i(t) = 10 + 5(\sin\ t)$ and the total EPS is the same as in scenario 3.

- **Scenario 7:**, As with scenario 6, but with the total EPS undergoing the more severe oscillations of case 4: $N_i(t) = s_i(10 + 7.5(\sin\ \pi t))/12$.

- **Scenario 8:** A structured population in which the size of each deme oscillates, according to the more severe version of scenarios 4 and 7, but such that the EPS of each deme is in exact sync with two demes (of differing size to itself) and exactly out of sync with the remaining three:

$$N_i(t) = \begin{cases} s_i(10 + 7.5(\sin\ \pi t))/12 & i \in \{1, 3, 5\} \\ N_i(10 + 7.5(\sin\ \pi(t + \pi)))/12 & i \in \{2, 4, 6\} \end{cases}$$

Note that $\sum_{i=1}^{6} N_i(t) = 10$; the total EPS is constant.

To convert the master phylogeny for each scenario to a set of sequences, the program πBUSS [14] was used. This works by placing a random, ancestral sequence at the root of the tree and letting it evolve along the tree's branches according to a stochastic model of sequence evolution. The sequence length and substitution process chosen was intended to roughly mimic the VP1 gene of FMDV; it had a length of 600bp and mutations occurred according to a strict molecular clock with a rate of 2.7×10^{-3} substitutions per site per year. Mutations occurred according to the HKY substitution model [61] with a transistion/transversion ratio of 2.718.

3.2.2 Subsampling for analysis

In every scenario, a variety of sampling schemes were used to select a subset of the master set for analysis. The 10-year sampling period was broken up into 40 intervals. An interval was picked according to a *temporal sampling scheme*. In unstructured scenarios, a sequence was picked (without replacement) from the subset of the master set whose sequence dates were in this interval uniformly at random. For structured scenarios, where every sample in this subset was also annotated with a deme, a sequence was picked a *spatial sampling scheme*. This was repeated until the desired number of samples was achieved.

Temporal sampling schemes explored were:

- **Uniform sampling:** All intervals have equal probability.
- **Proportional sampling:** Intervals are chosen with probability proportional to the value of the demographic function describing the total EPS, evaluated at the midpoint of the interval.
- **Reciprocal-proportional sampling:** Intervals are chosen with probability proportional to the reciprocal of value of that demographic function.

- **Exponential growth sampling:** Intervals are chosen with probability proportional to the value of the function $f(t) = e^{0.25t}$ evaluated at the midpoint of the interval. This was intended to represent a situation analogous to sampling at random from a public sequence database, as the number of sequences available from isolates taken in a given year will generally increase as the year increases.

Spatial sampling schemes explored were:

- **Uniform sampling:** All demes have equal probability.
- **Proportional sampling:** Demes were chosen with probability proportional to the EPS, relative to the EPSs of all other demes, of the deme at the midpoint of the interval.
- **Reciprocal-proportional sampling:** Demes are chosen with probability proportional to the reciprocal of the above.

In most cases, 300 samples were picked, and each sampling scheme was independently replicated 50 times. In some scenarios I also investigated the effect of varying the sample size; this was done by taking 5, or sometimes 10, replicates of sample sizes going from 25 to 500 in increments of 25 sequences.

An additional analysis was performed in scenario 5 only, in order to explore whether the population bottlenecks often seen towards the end of the timeline in skyline plots (as in chapter 2) could be the spurious result of an analysis that included many sequences acquired recently from a small geographical area. The sampling scheme for these was to randomly select 250 sequences using one of the above methods and then select an additional 50 at random from a single deme only during the last 0.25 years of the timeline. This was performed with each of the six demes in turn being the oversampled one.

3.2.3 MCMC analysis

The samples from each replicate of each sampling scheme was analysed separately in BEAST, assuming HKY as the nucleotide substitution model, a fixed molecular clock, and a skygrid tree prior [52]. The skygrid analysis had 199 grid points and a cut-off of 20 years, and unless otherwise stated the BEAST default Gamma$(0.001, 0.001)$ prior distribution was used on the precision parameter. In the first instance each MCMC chain was run for 30,000,000 states, sampling every 3,000 and discarding the first 10% as burn-in; all results were checked for an effective sample size (ESS) of at least 200 for all numerical model parameters and where this was not achieved, the burn-in was adjusted or the analysis re-run with a longer chain.

3.2.4 Performance evaluation

Skygrid reconstruction

The performance of the skygrid in reconstructing the demographic history of the simulated population was evaluated with three measures, two of which were used by Gill et al. [52] in their paper introducing the method. They are percent error, percent bias, and HPD size. As the behaviour of the reconstructed dynamics often diverged substantially and rapidly from reality in the period before sampling started (figure 3.2), I restricted my evaluation to the ten-year period during which sampling was taking place. Let R be the time of the last tip of the tree, in a timeline that goes from the start of sampling at $t = 0$ to its end at $t = 10$. Let $N(t)$ represent the true value of the EPS function at time t, $\hat{N}(t)$ the posterior median estimated EPS, $\hat{N}_{2.5}(t)$ the bottom of the 95% HPD interval and $\hat{N}_{97.5}(t)$

its top. The percent error is defined as:

$$100 \times \frac{1}{R} \int_0^R \frac{|\hat{N}(t) - N(t)|}{N(t)} dt$$

and represents the divergence of the median line of the reconstructed skygrid plot from the true curve of the EPS.

The percent bias is the same without the modulus:

$$100 \times \frac{1}{R} \int_0^R \frac{\hat{N}(t) - N(t)}{N(t)} dt$$

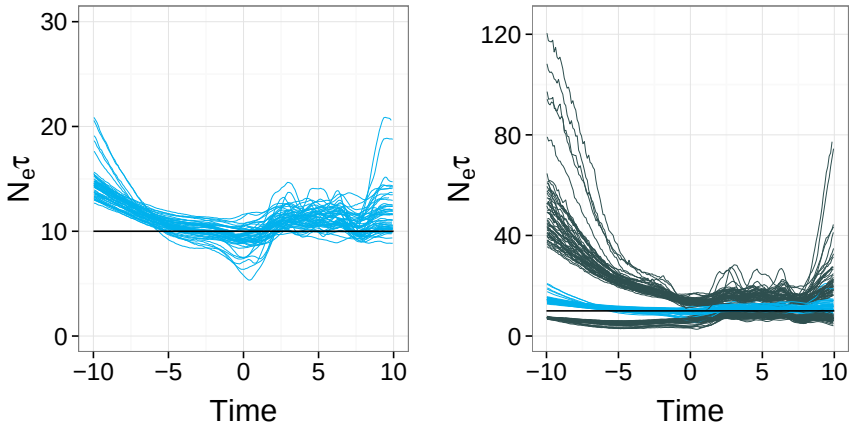
A negative value of this statistic represents a reconstruction in which the median line of the reconstruction is most often beneath the curve representing the true dynamics, a positive value represents one in which is it most often above it.

HPD size is a measure of the precision of the reconstruction:

$$\frac{1}{R} \int_0^R \frac{|\hat{N}_{97.5}(t) - \hat{N}_{2.5}(t)|}{N(t)} dt$$

with larger values reflecting wider credible intervals.

The values of these three statistics were calculated for the MCMC analysis that had been performed on every separate replicate of each sampling scheme. The results were then used as the basis for a kernel density estimate (KDE) for the probability density function of each statistic for each sampling scheme. These used a Gaussian kernel and bandwidth picked using the Sheather and Jones [133] method. Distributions were then compared by estimation of the coefficient of overlapping, using the OVL_5 estimator described by Schmid and Schmidt [130]. This estimates the area shared by both distributions, ranging from 0 if they have entirely disjoint support, and 1 if they are identical. Hypothesis tests were also employed to check whether features of the KDEs and hence coefficients of



(a) (b)

Figure 3.2: An illustration of the difference in the behaviour of reconstructions during the period while sampling was ongoing (times from 0 to 10) and the period before that. Blue lines are median lines from the reconstructed skygrid plot, from 50 replicates of uniform spatial and temporal sampling in scenario 1. The black line is the true total effective population size. Subfigure a) displays only the median lines while subfigure b) also displays lines marking the boundaries of the 95% HPD interval (grey).

overlapping were likely to be simply due to chance; as I felt it unwise to make assumptions about the distribution of these statistics, I used non-parametric tests. With so little prior research to base hypotheses on, a post-hoc testing strategy was employed, with the Nemenyi test identifying pairs of sampling schemes for which there was evidence that the distribution of each statistic was different. Test statistic calculations were conducted using the Tukey-Kramer method.

In some scenarios I investigated the relationship between percent error and sample size, and HPD size and sample size. The intention was to explore scenarios under which it would be prudent to acquire and analyse more samples. I used weighted least-squares regression [49] to fit curves of a variety of forms to this data. In some cases heteroscedasticity was an obvious feature of the output, so a model of

constant variance was not appropriate. The general form of these models for a statistic s of an analysis replicate of sample size n is $g(s) = Af(n) + B + \epsilon$, where A and B are constants and ϵ is normally distributed with mean 0 and variance $\sigma^2 v^2(n)$ where v is a positive function and σ^2 a scaling factor. A, B, σ^2, and the parameters of $v(n)$ are fit by the regression procedure. For g and f, I considered a linear relationship $s = An + B + \epsilon$, a logarithmic relationship $s = A\ln(n) + B + \epsilon$, a reciprocal relationship $s = \frac{A}{n} + B + \epsilon$, an exponential relationship $\ln(s) = An + B + \epsilon$, and a power law relationship $\ln(s) = A\ln(n) + B + \epsilon$. For the unscaled standard deviation $v(n)$ I considered a null model (such that the variance of estimates was not affected by sample size), an exponential relationship $v(n) = e^{tn}$, a power law relationship $v(n) = |n|^t$ and a power law plus constant $v(n) = t_1 + |n|^{t_2}$, where t, t_1 and t_2 are constants.

Modelled relationships were compared with each other using sample-size corrected Akaike information criterion (AICc); where the response variable was transformed, AICc values were corrected appropriately by the addition of the log of the Jacobian determinant of the transformation matrix. The model with the lowest AICc was taken to be the most appropriate.

Discrete traits phylogeography

The skygrid is a reconstruction of the temporal dynamics of the population. For a structured scenario, the spatial dynamics can also be explored. For the sake of simplicity, this was restricted to scenario 5; a structured population for which the EPS of every deme was time-invariant. The reconstruction was performed using the discrete-traits phylogeography model of Lemey et al. [94]. It should be noted that while the temporal dynamics were simulated under a coalescent model and reconstructed under a coalescent model, the skygrid, the spatial dynamics were simulated under a (structured) coalescent model and reconstructed under

the assumption that they evolved according to a CTMC, which is a very different mathematical process.

There are two options for reconstructing spatial dynamics using a CTMC discrete-traits model in BEAST. The first is almost exactly analogous to the GTR nucleotide substitution model [120] except that the Markov chain can have any number of states greater than 1. Discrete states, demes in this context, change based on a matrix of pairwise rates whose entries are never zero. The second method employs BSSVS, by which rates can also be zero, and a prior distribution is placed on the number of them that are. As the results of a MCMC analysis that employs BSSVS will give a posterior distribution for each entry in the rate matrix, with some values zero and some not, the hypothesis that the rate of movement between two states is in fact zero can be tested using Bayes Factors (BFs). This procedure is frequently used in phylogeography (as in chapter 2) to display lines on a map, representing links between two locations for which support for the hypothesis that the rate is nonzero reaches a certain level (typically BF=3). I analysed samples from the scenario 5 master set with both. I used a reversible rate model (assuming rates of transition between demes were equal in both directions, as was actually the case) and, as is the BEAST default for phylogeography, an equal-frequencies model. This makes the assumption that, over the long term, lineages will spend an equal amount of time in each deme.

Whether BSSVS is used or not, the results of a BEAST discrete traits analysis will give a posterior distribution for the rates of movement between each pair of states. With BSSVS enabled, however, these distributions are frequently bimodal, with many repeated values of 0 and the nonzero rates distributed according a peaked distribution. This makes taking a point estimate unwise, and thus when investigating numerical rate estimates I confined myself to the analysis that did not use the procedure. I used an estimate of the the maximum a posteriori (MAP) probability (in other words, the posterior mode) as the point estimate. Even in

this case, however, these rates are not directly comparable to the M_{ij} used to generate the simulated phylogenies as the former are with respect to forwards time and the latter backwards, so I calculated only a correlation coefficient (Kendall's τ) between the MAP estimates and the M_{ij}. As this statistic takes only a finite number of values for a given set of measurements (based on the number of ways that orderings can differ), its distribution over different sampling replicates is not continuous. Kernel density estimation, and hence calculation of the coefficient of overlapping, is not appropriate. Instead I used histogram intersection to compare distributions. I also again confirmed that differences in the distribution of Kendall's τ was not simply due to chance using Nemenyi post-hoc tests; in the latter case the chi-squared method was always used because, with a finite number of possible values for τ, ties are possible.

BSSVS, on the other hand, allows for a different type of investigation. The BF test is a binary classifier of zero or nonzero rates between each pair of states (or demes) in the analysis. In a simulated scenario such as this one, in which the truth is known, this can be used to calculate overall accuracy, sensitivity and specificity for a given BF value, or to draw a receiver-operator curve (ROC) evaluating the performance of the classifier. As with Kendall's τ, these statistics take only a finite set of values for a given set of measurements. I again compared these statistics across sampling schemes using histogram intersection and chi-squared version of Nemenyi post-hoc tests.

The relationship between τ and sample size for the non-BSSVS analysis, and between the accuracy of BSSVS as a classifier and sample size for the BSSVS analysis, were also investigated using the same weighted least-squares regression procedure outlined above.

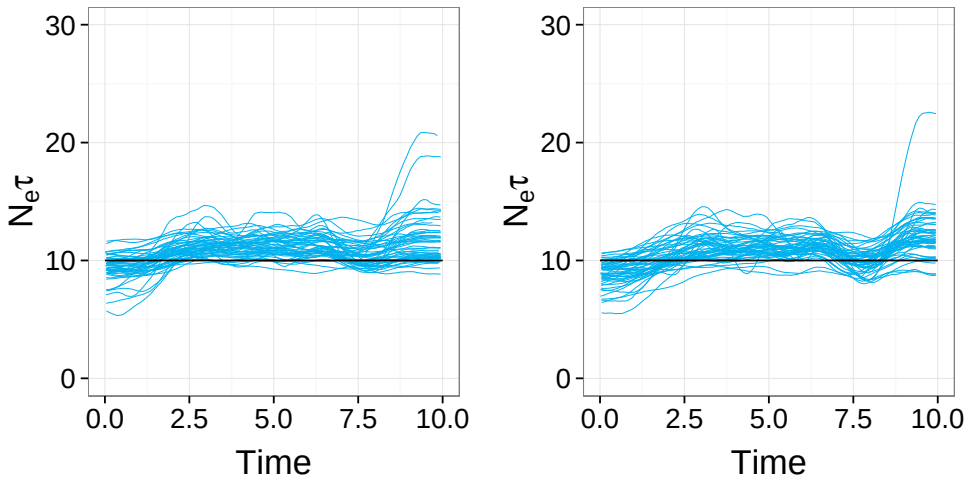
3.3 Results

3.3.1 Skygrid reconstruction

Scenario 1: Single population, constant size

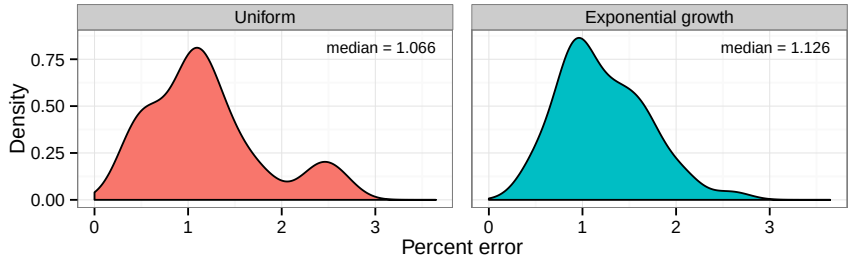
Figure 3.3 overlays the median lines for 50 reconstructed skygrid plots reconstructed from analyses with samples chosen by uniform and exponential growth temporal sampling, whereas figure 3.4 displays KDEs for the distribution of the percent error, percent bias and HPD size statistics. A bias towards overestimating population sizes is clear. Coefficient of overlapping estimates were 0.7 for percent error, 0.76 for percent bias and 0.66 for HPD size. While the KDE graphs and medians might suggest superior performance for the exponential growth scheme in terms of percent bias, there was little evidence of this ($p = 0.154$); nor was there any suggestion of a superior performance for either in terms of percent error (p=0.563). However, there was a suggestion that the reconstruction was more precise with uniform sampling ($p = 0.0136$). As even in this simple situation there was evidence that letting the number of sequences increase with time performed worse than, effectively, stratifying by sampling period, I do not consider the former further in other scenarios.

It is apparent that the performance of the skygrid method in reconstructing the true dynamics is variable, even when the samples are chosen according to different replicates of the same scheme. For example, figure 3.5 shows, for the uniform sampling scheme, all reconstructed plots for the 50 replicates sorted in order of increasing percent error. The best reconstructions are nearly flawless, whereas the worst have spurious features that might lead an unwary researcher to the wrong conclusions. However, the line representing the true EPS does lie within the 95% HPDI for the entire length of the sampling period in the considerable majority

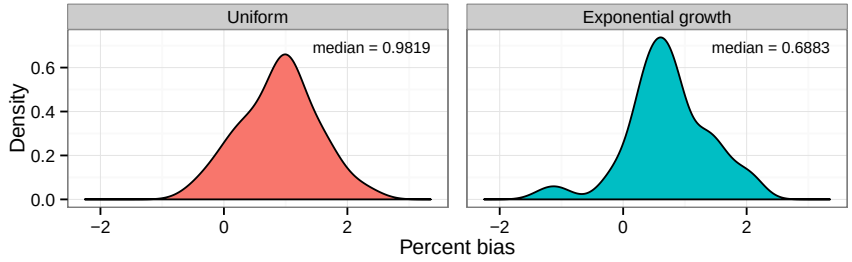


(a) Uniform **(b)** Exponential growth

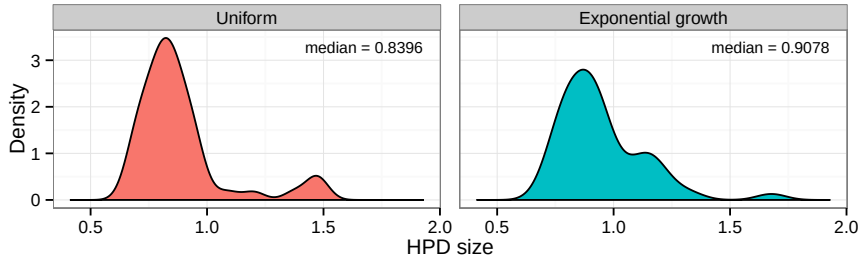
Figure 3.3: Overlaid median lines for 50 reconstructed skygrid plots for different sampling schemes in scenario 1: a) exponential increase in probability of selection over time, b) uniform probability of selection over time. The graph is limited to the 10 time units during which sampling was taking place. The red line is the true effective population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.4: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 1: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

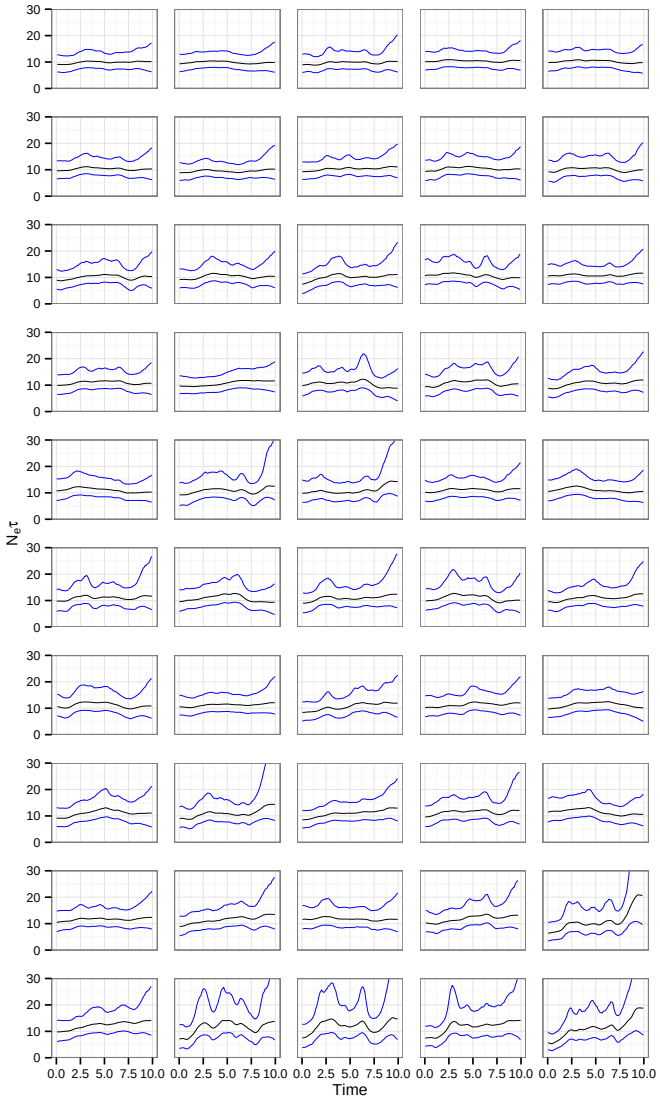


Figure 3.5: Skygrid reconstructions for the 50 replicates of the uniform sampling scheme in scenario 1, sorted by increasing percent error. The black line is the median estimate, the blue lines the bounds of the 95% HPD interval.

of replicates. To further explore this, I performed a model comparison between the skygrid (which allows for population dynamics that are not governed by any deterministic function) and a model of constant population size (which should, in this case, be sufficient as it represents the true dynamics). The sequences from the five replicates with the highest percent error (the bottom row of figure 3.5) were re-analysed, replacing the skygrid tree prior with the constant model, and both models were compared by calculating marginal likelihood estimates (MLEs) using both path-sampling (PS) and stepping-stone sampling (SS) [6]. Ratios of the MLEs were calculated to give a BF comparing the two models.

The results of this are summarised in table 3.1. BFs greater than 1 support the skygrid over the constant model. In two cases (replicate IDs 12 and 50), the constant model is favoured by both estimation methods. In one (ID 1), it is favoured by SS but the skygrid is slightly preferred by PS; nevertheless the BF is only slightly greater than 1 and this would not be interpreted as conclusive. However, two replicates, IDs 22 and 43, give figures that would support the rejection of a constant size population model in favour of more complex dynamics. These dynamics are purely a sampling artefact. In particular, for replicate 43 the difference is dramatic and the hypothesis of constant size would, with no knowledge of the true situation, be decisively rejected.

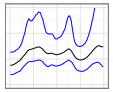
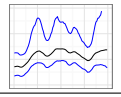
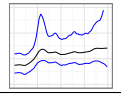
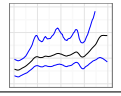
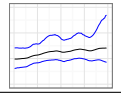
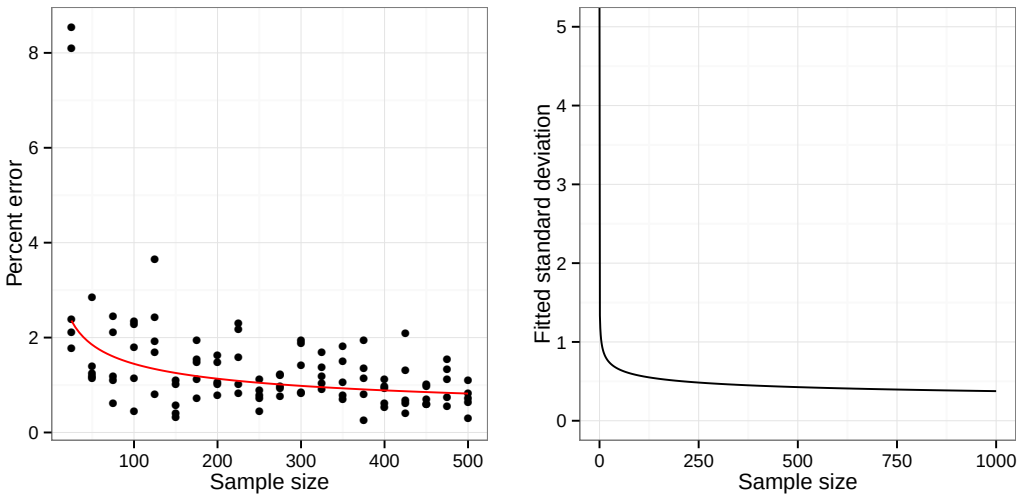
ID	Skygrid graph	Path-sampling			Stepping-stone sampling		
		log MLE (Constant)	log MLE (Skygrid)	BF	log MLE (Constant)	log MLE (Skygrid)	BF
1		-5665.63	-5655.55	1.08	-5655.48	-5656.52	0.35
12		-5780.95	-5782.16	0.30	-5782.33	-5784.48	0.12
22		-5783.08	-5782.00	2.92	-5785.65	-5782.94	15.00
43		-5897.60	-5892.42	176.81	-5899.30	-5893.56	310.21
50		-6012.19	-6012.22	0.97	-6012.63	-6014.24	0.20

Table 3.1: Results of marginal likelihood estimation. The five replicates of the uniform sampling scheme, scenario 1, whose reconstructed skygrid plots had highest percent error in the median line were re-analysed using both the skygrid and a constant population size coalescent model as tree priors. Figures given are the log marginal likelihoods for both priors, estimated using both path-sampling and stepping-stone sampling. The BFs given are for the hypothesis that the skygrid model fits the data better than the constant population model.

To test the results of variation in sample size, uniform temporal sampling was used to select datasets of sizes from 25 to 500 in increments of 25, with 5 replicates for each. Figures 3.6a and 3.7a plot the results of this as percent error and HPD size statistics against sample size. As might be expected, some heteroscedasticity is evident, with the variance of the statistics for the set of replicates of the same sample size decreasing as that sample size increases.

Table 3.2 gives AICc values for models of the relationship between percent error and sample size, and table 3.3 does the same for HPD size and sample size. For both statistics a power law model is preferred, although there is often little to pick between different variance models. For percent error the variance model with lowest AICc was a power law $v(n) = |n|^t$ and the best-fitted values are $A = -0.35$, $B = 1.99$, $\sigma^2 = 1.78$ and $t = -0.18$. For HPD size the model was a power plus constant $v(n) = t_1 + |n|^{t_2}$ andd the best fitted values are $A = -0.47$, $B = 2.53$, $\sigma^2 = 477.98$, $t_1 = 5.11 \times 10^{-3}$ and $t_2 = -1.11$. The curves representing the models with the lowest AICc values and their fitted parameter values have been added to figures 3.6a and 3.7a in red, and figures 3.6b and 3.7b are the curves for the corresponding standard deviation models.

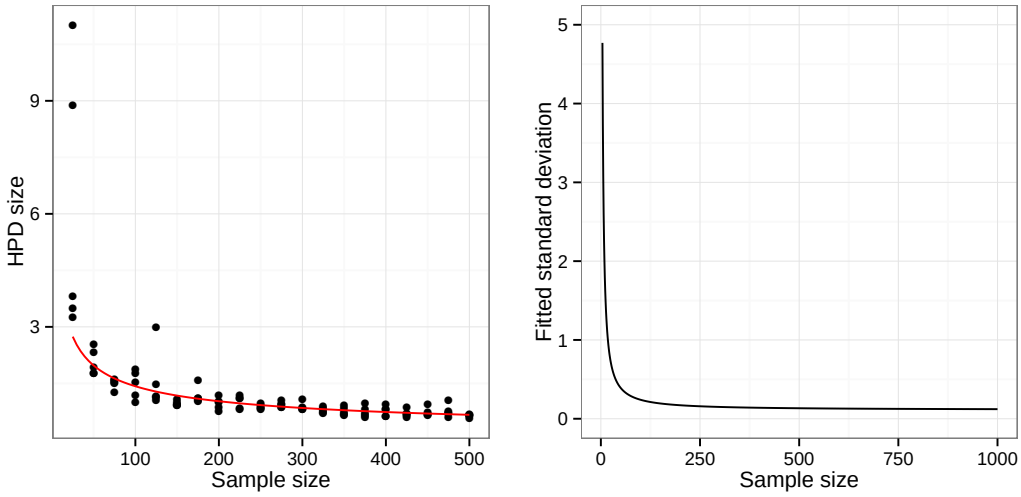


(a) **(b)**

Figure 3.6: a) Scatter plot of percent error versus sample size for 100 replicates of the uniform sampling scheme in scenario 1. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. b) Curve of the standard deviation function of the best-fit model.

Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
$\text{error} = An + B$	317.42	234.42	210.90	209.23
$\text{error} = A\ln(n) + B$	292.50	218.05	198.08	197.17
$\text{error} = \frac{A}{n} + B$	267.45	203.38	186.82	186.87
$\ln(\text{error}) = An + B$	202.30	198.12	195.72	197.40
$\ln(\text{error}) = A\ln(n) + B$	184.88	183.34	182.43	184.62

Table 3.2: AICc values for models of the relationship between percent error and sample size, scenario 1, whose parameters were fit by least-squares regression.



(a) **(b)**

Figure 3.7: a) Scatter plot of HPD size versus sample size for 100 replicates of the uniform sampling scheme in scenario 1. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. b) Curve of the standard deviation function of the best-fit model.

Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
$\text{size} = An + B$	340.26	107.58	34.11	28.59
$\text{size} = A\ln(n) + B$	297.42	84.44	12.58	6.79
$\text{size} = \frac{A}{n} + B$	243.74	44.66	-8.75	-10.53
$\ln(\text{size}) = An + B$	95.91	36.60	7.89	2.83
$\ln(\text{size}) = A\ln(n) + B$	18.82	-13.21	-26.58	-27.08

Table 3.3: AICc values for models of the relationship between HPD size and sample size, scenario 1, whose parameters were fit by least-squares regression.

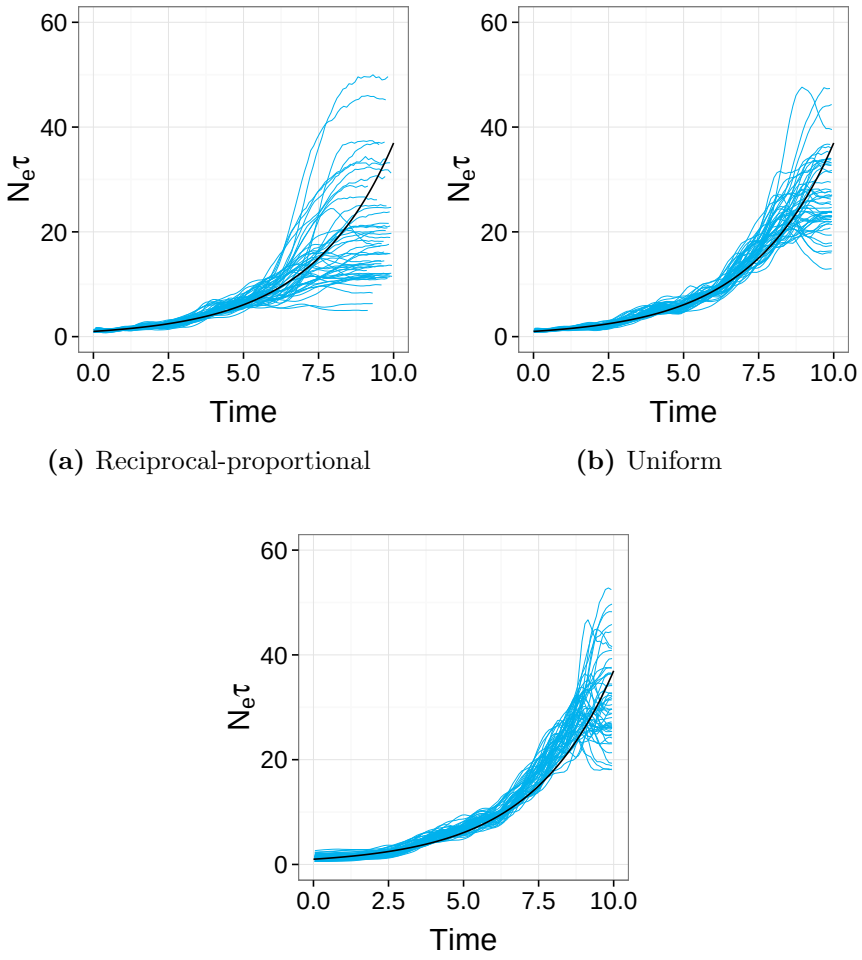
Scenario 2: Single population, exponential growth

Figure 3.8 shows the overlaid median lines for 50 reconstructed plots for uniform, proportional, and reciprocal-proportional sampling. It can be seen that, as previously observed [31], a spurious flattening out or decline in the median line towards the end of sampling is common. This is not prevented by any particular sampling scheme, but some are more prone to earlier departures from the growth curve representing the true dynamics than others. Figure 3.9 displays KDEs for percent error, percent bias, and HPD size. Coefficient of overlapping estimates are given in table 3.4, along with p-values from post-hoc tests. There is considerable disagreement between the distribution for the uniform scheme and the other two, with the former being more accurate and precise, for percent error and HPD size. On the other hand, the reciprocal-proportional scheme shows the least bias. (While figure 3.8 would, at first glance, suggest much larger error for the reciprocal-proportional scheme than any other, the percent error statistic scales the error at a certain time point by the true value of the EPS at that point, and the median line for the proportional scheme generally deviates much more from the true line in the early part of the timeline than the other two do.)

Scenario 3: Single population, long-period oscillations

The overlaid median lines for uniform, proportional and reciprocal-proportional temporal sampling can be seen in figure 3.10. Of note, while all three schemes have a bias towards overestimating EPSs when the true value is at its minimum, the effect is smallest for reciprocal-proportional sampling. The KDE plots (figure 3.11) also suggest that reciprocal-proportional sampling is preferable, and estimated coefficients of overlapping (table 3.5), combined with post-hoc test results, show a particular superiority of this over proportional sampling for every statistic,

although its advantage over uniform sampling in terms of error and HPD size may be due to chance.

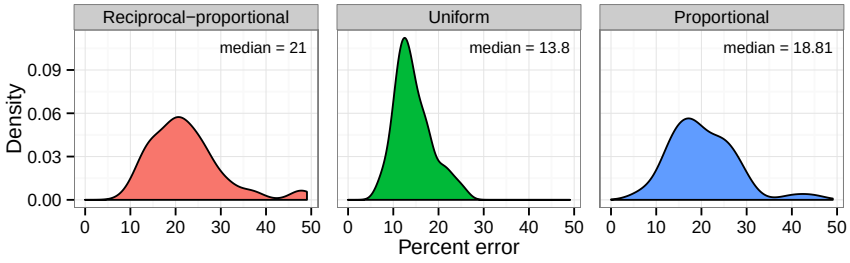


(a) Reciprocal-proportional

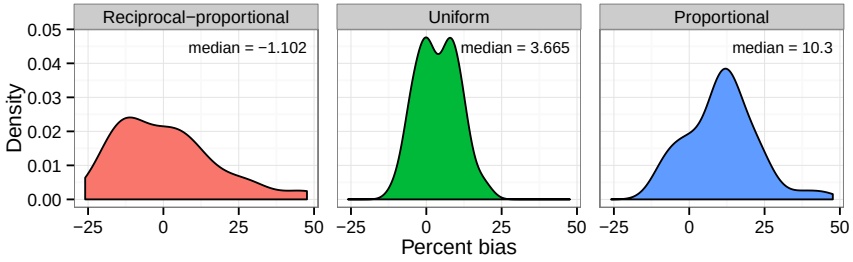
(b) Uniform

(c) Proportional

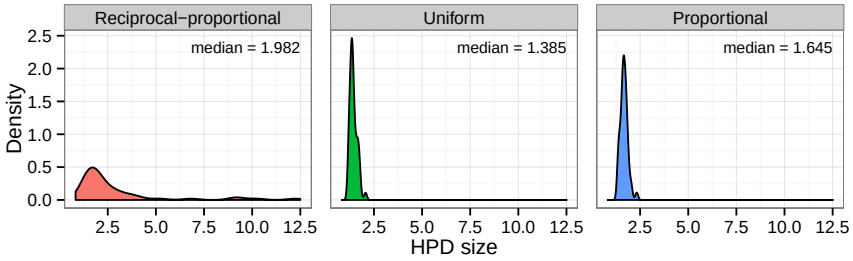
Figure 3.8: Overlaid median lines for 50 reconstructed skygrid plots for scenario 2: a) probability of inclusion proportional to the reciprocal of effective population size, b) uniform probability of inclusion, c) probability of inclusion proportional to effective population size. The red line is the true effective population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.9: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 2: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

	Reciprocal-proportional	Uniform
Uniform	**0.5** (2.15×10^{-8})	
Proportional	0.88 (0.494)	**0.62** (9.71×10^{-6})

(a) Percent error

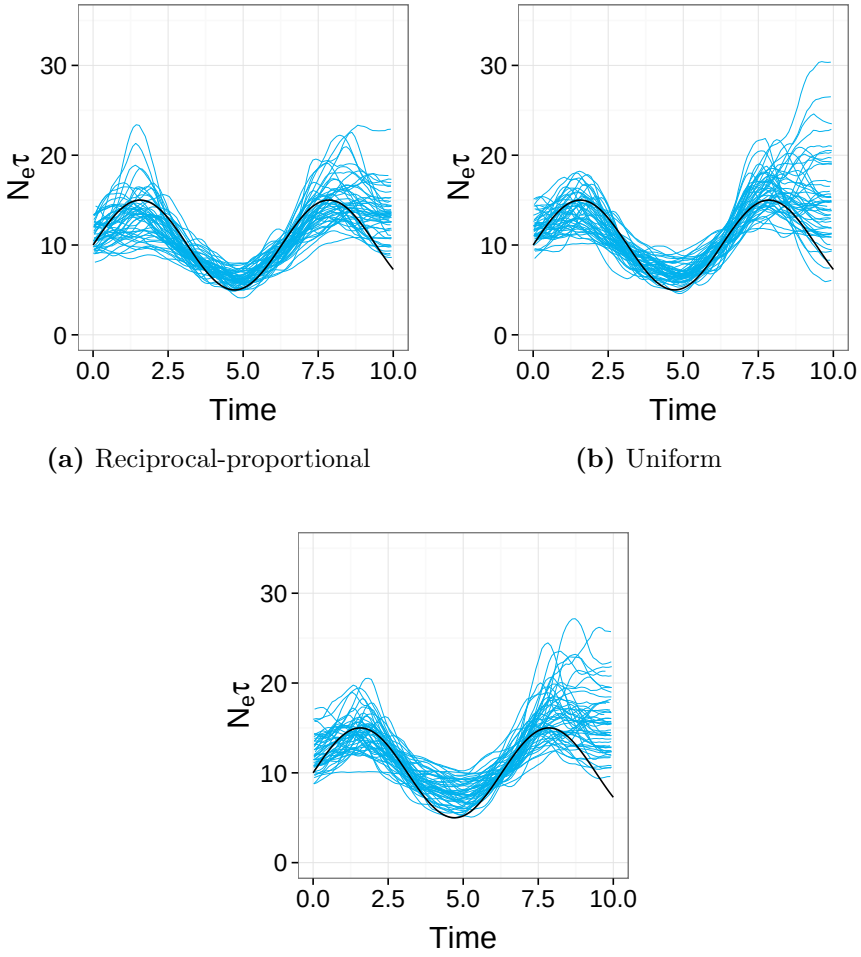
	Reciprocal-proportional	Uniform
Uniform	0.5 (0.367)	
Proportional	**0.66** (5.46×10^{5})	**0.7** (9.43×10^{-3})

(b) Percent bias

	Reciprocal-proportional	Uniform
Uniform	**0.44** (1.77×10^{-11})	
Proportional	0.48 (0.0509)	**0.5** (1.63×10^{-5})

(c) HPD size

Table 3.4: Estimated coefficients of overlapping for distributions of statistics in scenario 2. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.

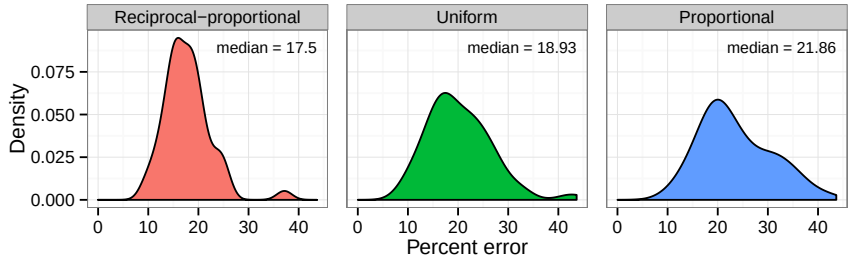


(a) Reciprocal-proportional

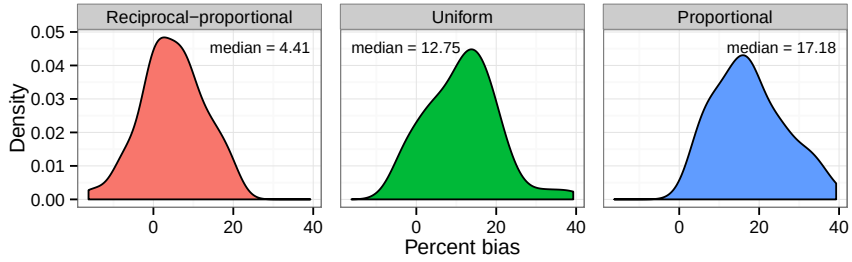
(b) Uniform

(c) Proportional

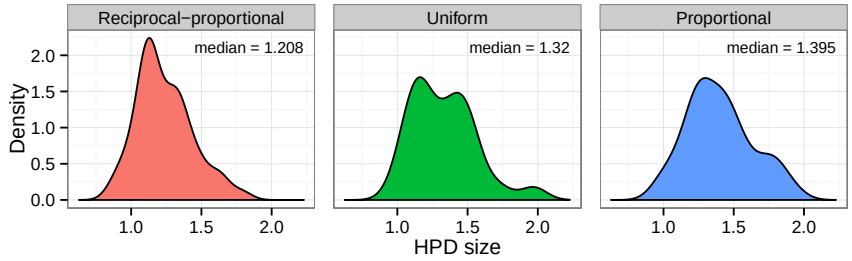
Figure 3.10: Overlaid median lines for 50 reconstructed skygrid plots for scenario 3: a) probability of inclusion proportional to the reciprocal of effective population size, b) uniform probability of inclusion, c) probability of inclusion proportional to effective population size. The red line is the true effective population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.11: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 3: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

	Reciprocal-proportional	Uniform
Uniform	0.72 (0.0872)	
Proportional	**0.64** (2.14×10^{-5})	**0.82** (0.0461)

(a) Percent error

	Reciprocal-proportional	Uniform
Uniform	**0.64** (2.12×10^{-3})	
Proportional	**0.44** (5.3×10^{-10})	**0.76** (7.52×10^{-3})

(b) Percent bias

	Reciprocal-proportional	Uniform
Uniform	0.78 (0.107)	
Proportional	**0.74** (1×10^{-3})	0.72 (0.264)

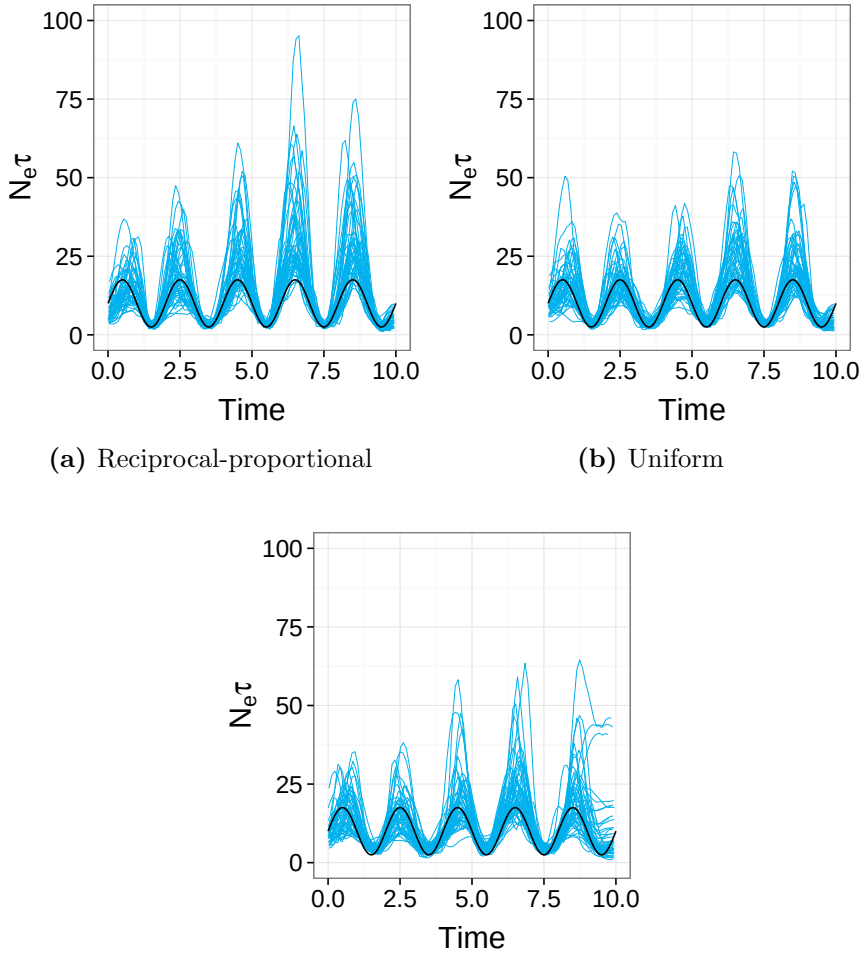
(c) HPD size

Table 3.5: Estimated coefficients of overlapping for distributions of statistics in scenario 3. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.

Scenario 4: Single population, short-period oscillations

Figure 3.12 shows the overlaid plots. Once again, reciprocal-proportional sampling is most effective at capturing the dynamics when the EPS is at a minimum, but in this case this is balanced by a tendency to overestimate sizes at the maximum by greater amounts than the other two schemes. The KDE plots (figure 3.13) suggest that the relative weaknesses of proportional and reciprocal-proportional sampling cancel each other out in this case, leaving uniform sampling as the best-performing scheme. There is more overlap in KDEs here than was seen in scenario 3 (table 3.6), with post-hoc tests only suggesting evidence for the superiority of uniform over proportional sampling for error and bias, and over both the other schemes for HPD size.

I repeated the investigation of the effect of sample size from scenario 1 for this scenario. Scatter plots for sample size against percent error and HPD size are figures 3.15a and 3.16a. Figure 3.14 displays the reconstructed plots for every replicate. Some extreme outliers are omitted from the scatter plots. One replicate with a sample size of 25 (the rightmost graph in the first row of figure 3.14) has a percent error of 890.11 and an HPD size of 1.88×10^8; the HPD region becomes extremely wide in the second half of the sampling interval, with the upper limit peaking at 4.42×10^9. Repeated BEAST runs on the same dataset did not change this behaviour. Another replicate with a sample size of 100 (second from left, fourth row) also showed a wide HPD interval at the very end of the timeline (HPD size = 1333.46). In addition, a reduction in HPD size for low sample sizes is apparent, and the reason for this can be seen in figure 3.14: with fewer samples the oscillations tend to be "damped", giving a median line that suggests constant dynamics, and a much narrower HPD interval. I will return to this subject in more detail when discussing scenario 7. For this scenario, I simply excluded all replicates for which the oscillations were so damped that, when examined by

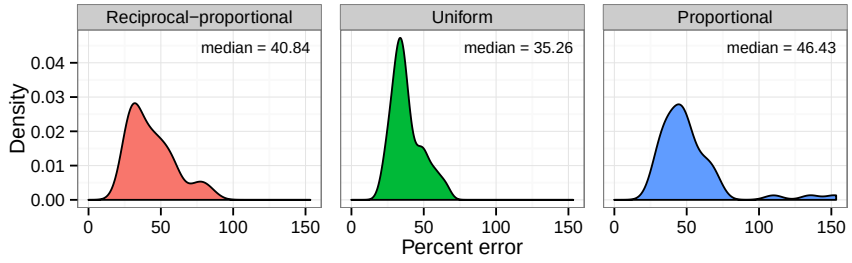


(a) Reciprocal-proportional

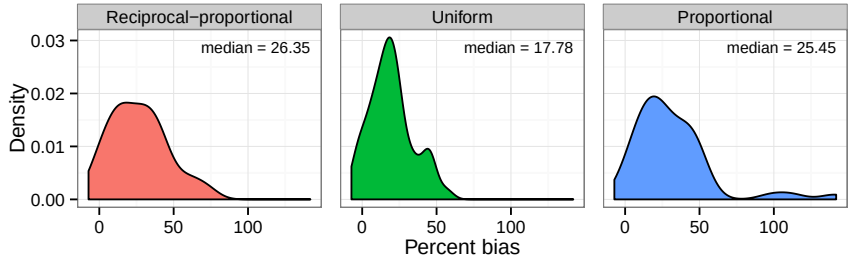
(b) Uniform

(c) Proportional

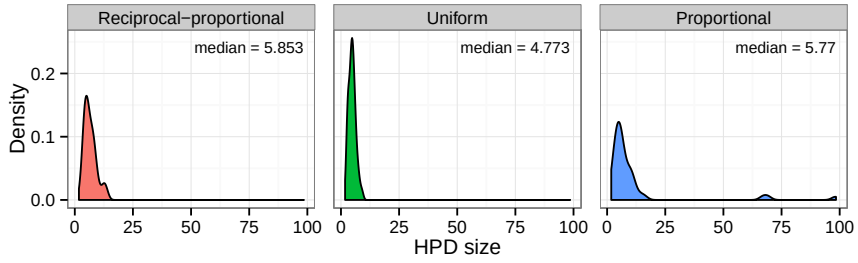
Figure 3.12: Overlaid median lines for 50 reconstructed skygrid plots for scenario 4: a) probability of inclusion proportional to the reciprocal of effective population size, b) uniform probability of inclusion, c) probability of inclusion proportional to effective population size. The red line is the true effective population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.13: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 4: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

	Reciprocal-proportional	Uniform
Uniform	0.8 (0.289)	
Proportional	0.7 (0.0699)	**0.62** (6.06×10^{-4})

(a) Percent error

	Reciprocal-proportional	Uniform
Uniform	0.76 (0.166)	
Proportional	0.76 (0.807)	**0.76** (0.0395)

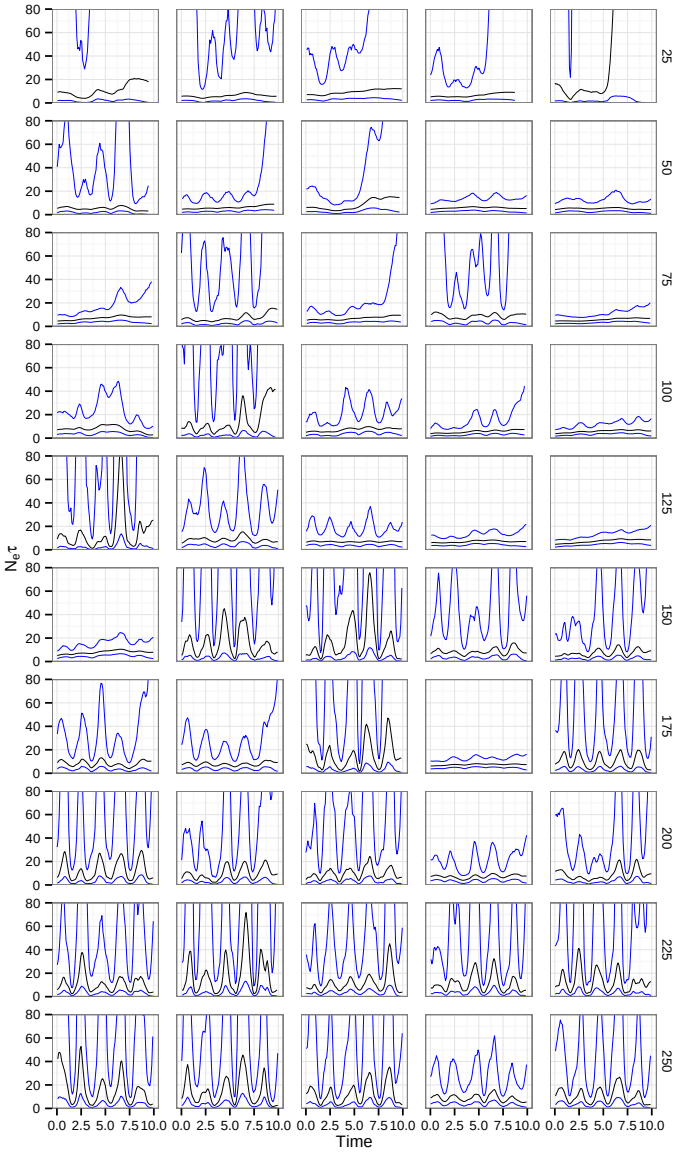
(b) Percent bias

	Reciprocal-proportional	Uniform
Uniform	**0.68** (2.12×10^{-3})	
Proportional	0.72 (0.984)	**0.7** (3.87×10^{-3})

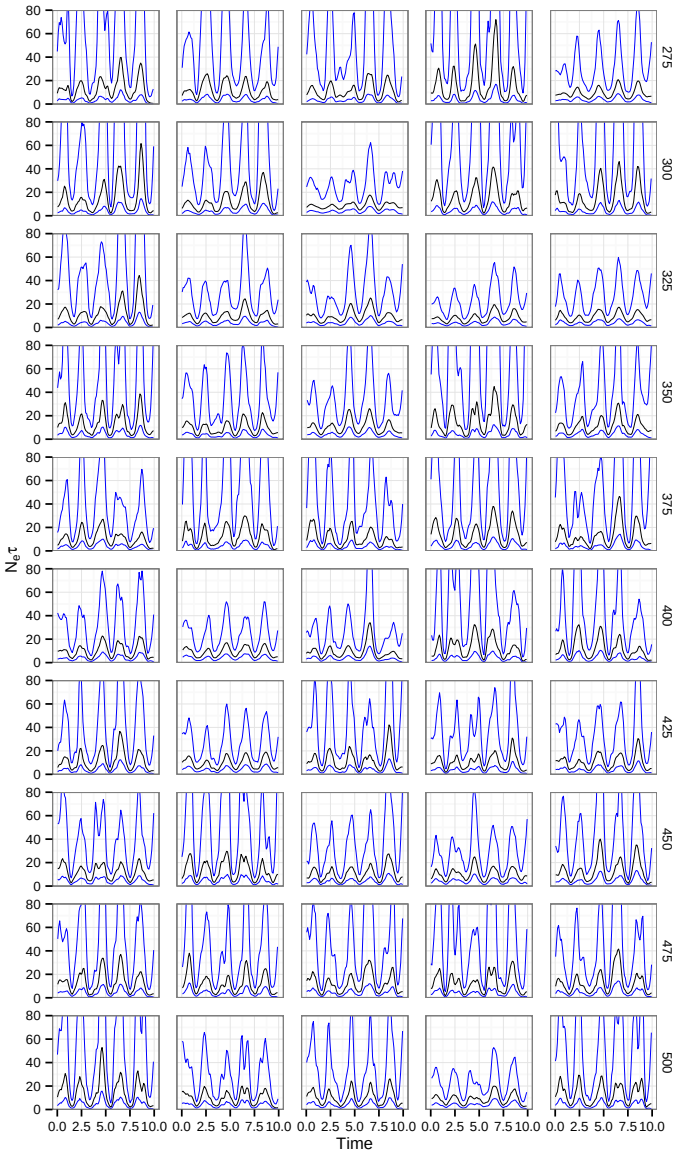
(c) HPD size

Table 3.6: Estimated coefficients of overlapping for distributions of statistics in scenario 4. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.

eye, there was no suggestion of oscillatory behaviour in the median line at all. These correspond to points in red on the scatter plot; the first outlier described above was excluded, but the second was not. I once again used weighted least squares regression to model the relationship between sample size, percent error and HPD size, with the replicates that showed damping removed. The curves representing the best-fit model are superimposed on figures 3.15a and 3.16a, and the AICc scores are given in tables 3.7 and 3.8. Reciprocal plus constant models were preferred for both. For percent error the variance model with lowest AICc was a power law $v(n) = |n|^t$ and the best-fitted values are $A = 6019.65$, $B = 20.56$, $\sigma^2 = 8.10 \times 10^5$ and $t = -0.78$. For HPD size the preferred model was a power plus constant $v(n) = t_1 + |n|^{t_2}$ and the best fitted values are $A = -1343.66$, $B = 0.64$, $\sigma^2 = 2.46 \times 10^{37}$, $t_1 = 2.99 \times 10^{-19}$ and $t_2 = -7.98$. The functions representing the standard deviation of the error in the best-fit model are shown in figures 3.15b and 3.16b.

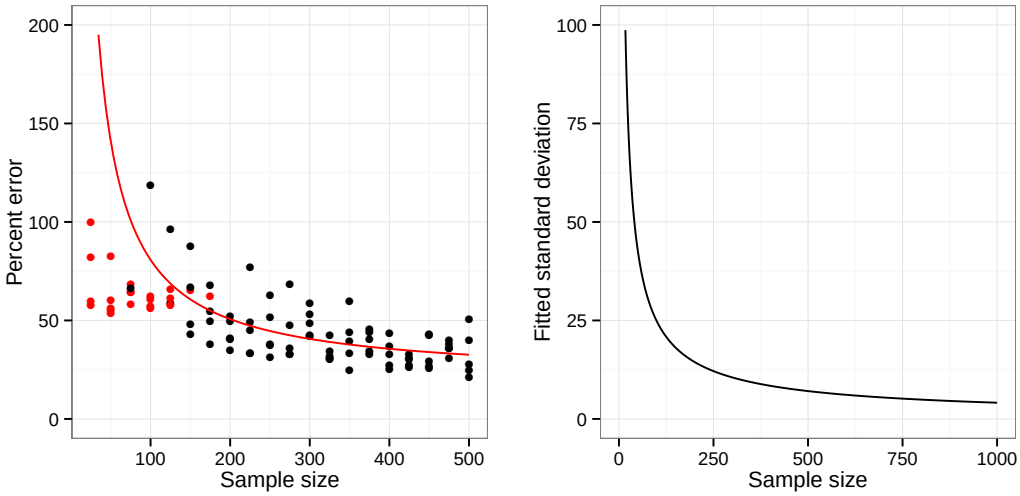


(a)



(b)

Figure 3.14: Reconstructed skygrid plots for scenario 4. Each row is 5 replicates of the same sample size, given on the right. The black line is the median estimate, the blue lines the bounds of the 95% HPD interval. Subfigures: a) Sample sizes 25-250. b) Sample sizes 275-500.

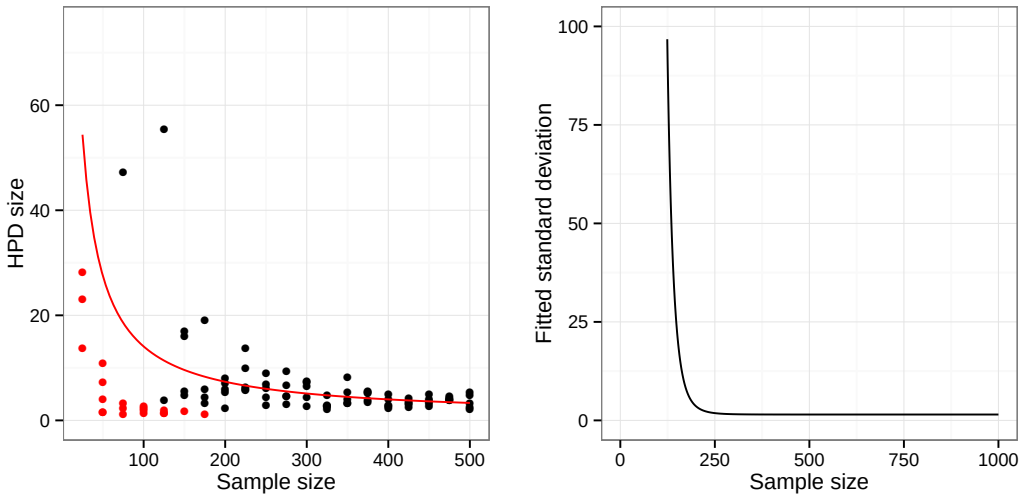


(a) **(b)**

Figure 3.15: a) Scatter plot of percent error versus sample size for 100 replicates of the uniform sampling scene in scenario 4. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. The red points were replicates that displayed "damped" oscillations and were not used to fit the model. One extreme outlier (which was not fit to) is not shown. b) Curve of the standard deviation function of the best-fit model.

Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
$\text{error} = An + B$	626.09	606.19	601.70	602.11
$\text{error} = A\ln(n) + B$	605.92	589.11	586.06	587.51
$\text{error} = \frac{A}{n} + B$	593.30	575.46	572.49	574.23
$\ln(\text{error}) = An + B$	604.43	604.84	603.98	605.46
$\ln(\text{error}) = A\ln(n) + B$	587.65	588.77	588.36	590.44

Table 3.7: AICc values for models of the relationship between percent error and sample size, scenario 4, whose parameters were fit by least-squares regression.



(a) **(b)**

Figure 3.16: a) Scatter plot of HPD size versus sample size for 100 replicates of the uniform sampling scene in scenario 4. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. The red points were replicates that displayed "damped" oscillations and were not used to fit the model. Two extreme outliers, one of which was fit to and one of which was not (see the text) are not shown. b) Curve of the standard deviation function of the best-fit model.

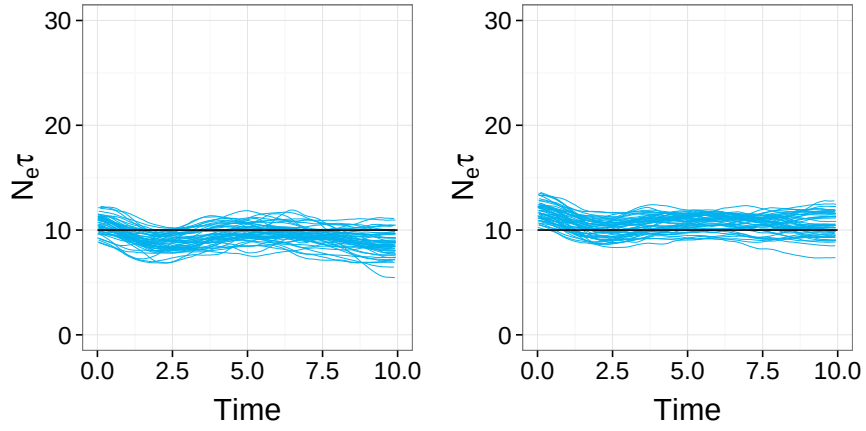
Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
$\text{size} = An + B$	986.81	559.25	462.38	400.86
$\text{size} = A\ln(n) + B$	971.95	547.02	450.04	386.98
$\text{size} = \frac{A}{n} + B$	956.50	534.74	437.65	373.83
$\ln(\text{size}) = An + B$	562.29	501.00	488.82	484.80
$\ln(\text{size}) = A\ln(n) + B$	539.38	484.51	473.76	470.68

Table 3.8: AICc values for models of the relationship between HPD size and sample size, scenario 4, whose parameters were fit by least-squares regression.

Scenario 5: Structured population, constant size

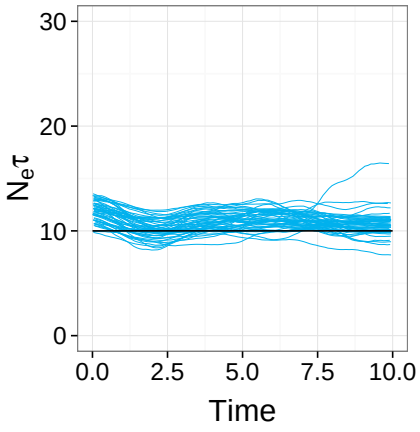
Figure 3.17 gives the overlaid plots, and figure 3.18 the KDEs. Note that in this case, the proportionality or reciprocal-proportionality refers to spatial sampling; as the overall population size was constant, I used uniform temporal sampling only here. For coefficients of overlapping and results of post-hoc tests, see table 3.9. The performance of the uniform and proportional schemes are basically equivalent, but reciprocal-proportional sampling is very different: it is no more accurate, but the bias occurs in the opposite direction (as is clear in figure 3.18b) and it also gives slightly more precise reconstructions.

This is the scenario in which the effect of oversampling a single deme towards the end of the timeline was investigated. (The 250 sequences in these analyses that were not part of the oversampling were selected using reciprocal-proportional spatial sampling, as this was marginally the best-performing scheme.) When overlaying the plots (figure 3.19) a spurious bottleneck effect is immediately clear, and it is more extreme if the oversampled deme is smaller. The true value of $N_e\tau$ was outside the 95% HPD interval at the very end of the timeline in 100% of replicates where the oversampled deme was small, 98% where it was medium-sized, and 60% where it was large.



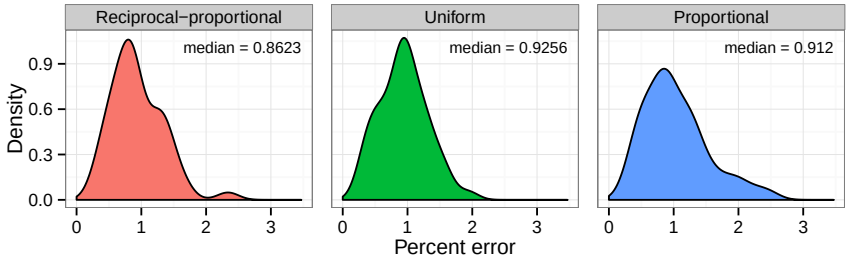
(a) Reciprocal-proportional

(b) Uniform

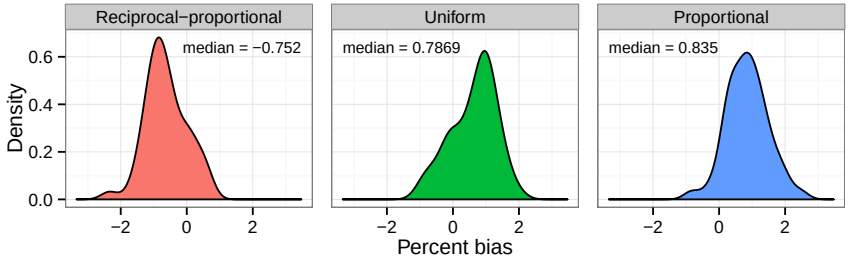


(c) Proportional

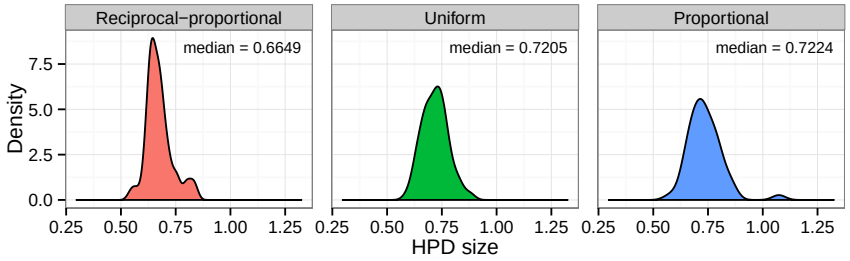
Figure 3.17: Overlaid median lines for 50 reconstructed skygrid plots for scenario 5: a) probability of inclusion proportional to the reciprocal of effective population size, b) uniform probability of inclusion, c) probability of inclusion proportional to effective population size. The red line is the true effective population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.18: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 5: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

	Reciprocal-proportional	Uniform
Uniform	0.78 (0.935)	
Proportional	0.84 (0.671)	0.98 (0.87)

(a) Percent error

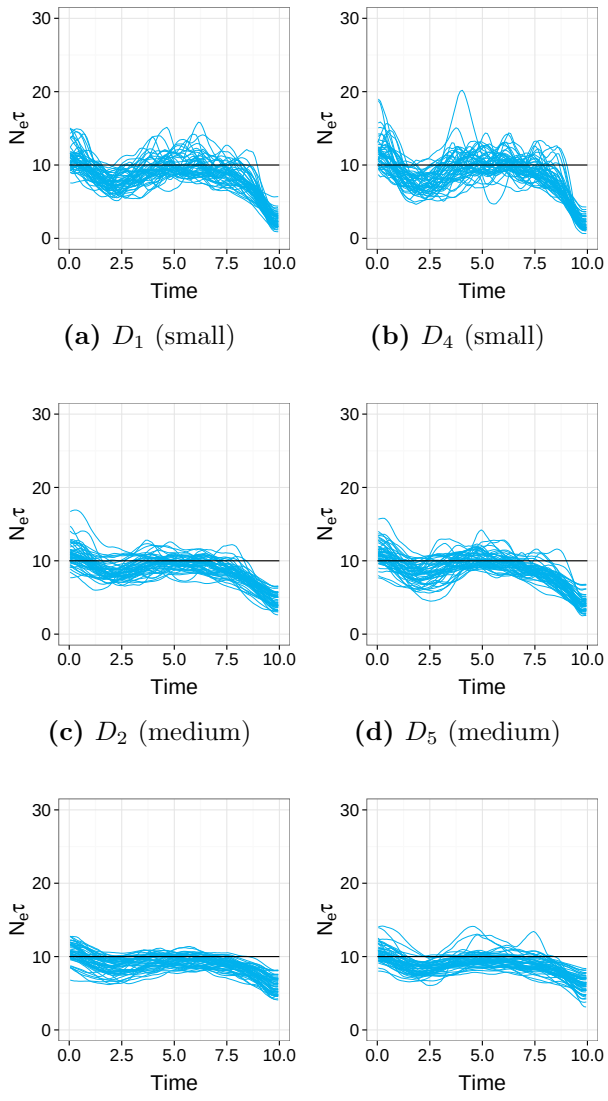
	Reciprocal-proportional	Uniform
Uniform	**0.44** (2.6×10^{-11})	
Proportional	**0.22** (4.8×10^{-14})	0.76 (0.504)

(b) Percent bias

	Reciprocal-proportional	Uniform
Uniform	**0.64** (4.61×10^{-4})	
Proportional	**0.54** (9.18×10^{-6})	0.9 (0.649)

(c) HPD size

Table 3.9: Estimated coefficients of overlapping for distributions of statistics in scenario 5. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.



(a) D_1 (small) **(b)** D_4 (small)

(c) D_2 (medium) **(d)** D_5 (medium)

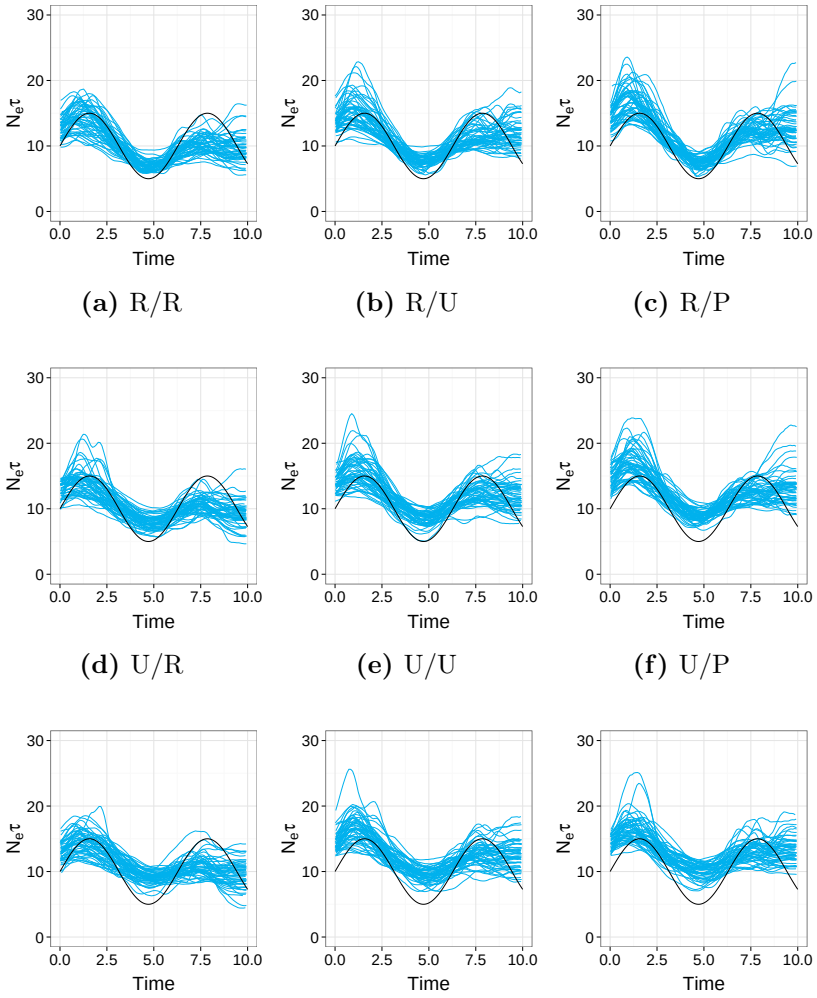
(e) D_3 (large) **(f)** D_6 (large)

Figure 3.19: Overlaid median lines for 50 reconstructed skygrid plots for scenario 5, where additional samples are selected from one deme in the last 0.25 years of the timeline. The red line is the true population size. a) and b) D_1 and D_4 (small). c) and d) D_2 and D_5 (medium). e) and f) D_3 and D_6 (large).

Scenario 6: Structured population, long-period oscillations

Sampling schemes now have two components: a temporal scheme and a spatial scheme. There are nine possible combinations. The overlaid median plots can be seen in figure 3.20, and the KDEs in figure 3.21. Notably, moving from proportional to uniform sampling, and from uniform to reciprocal-proportional, decreases the median percent error and percent bias for both space and time. The same is true for HPD size with respect to the spatial scheme, but the difference is much more modest.

For the coefficients of overlapping and p-values for post-hoc tests, see table 3.10. (The estimated coefficient of overlapping of 1.02 in table 3.10c when comparing proportional and reciprocal-proportional temporal schemes with the reciprocal-proportional spatial scheme is the result of a flaw in the estimator when comparing very similar distributions.) Regardless of the spatial scheme, the advantage in terms of error and bias in moving from proportional to reciprocal-proportional temporal sampling (figures in red in table 3.10) is very clear, but this does not apply to HPD size, for which variation could easily be due to chance; this is quite analogous to scenario 3. When the temporal scheme is fixed, the equivalent comparison (blue in the table) shows much more overlap in the KDEs for error, but in the distributions for bias the coefficient is never greater than 0.34 and there is also less overlap for HPD size; this is similar to scenario 5.

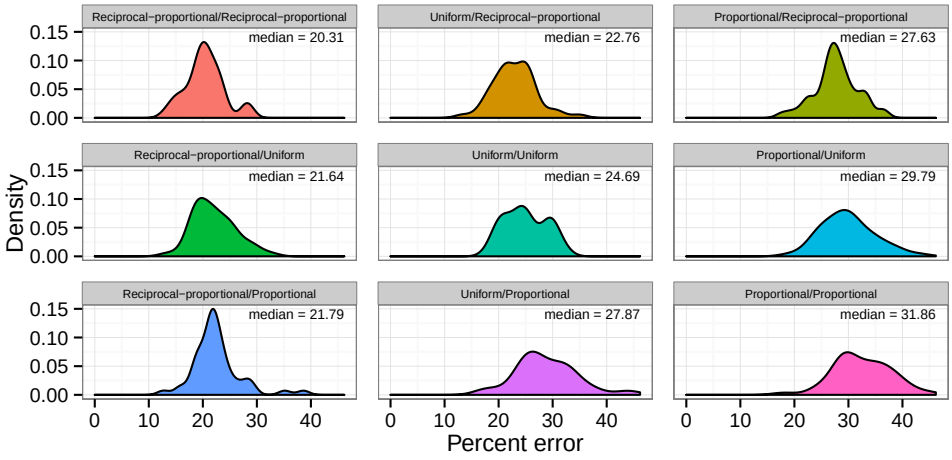


(a) R/R **(b)** R/U **(c)** R/P

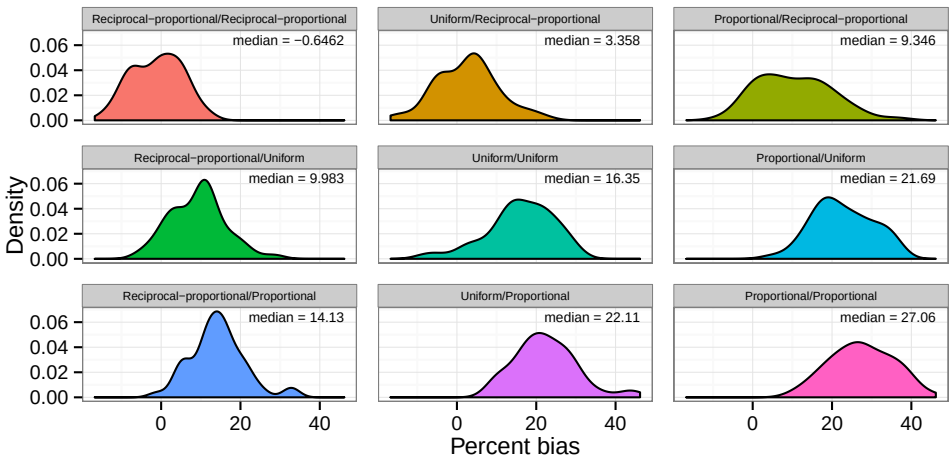
(d) U/R **(e)** U/U **(f)** U/P

(g) P/R **(h)** P/U **(i)** P/P

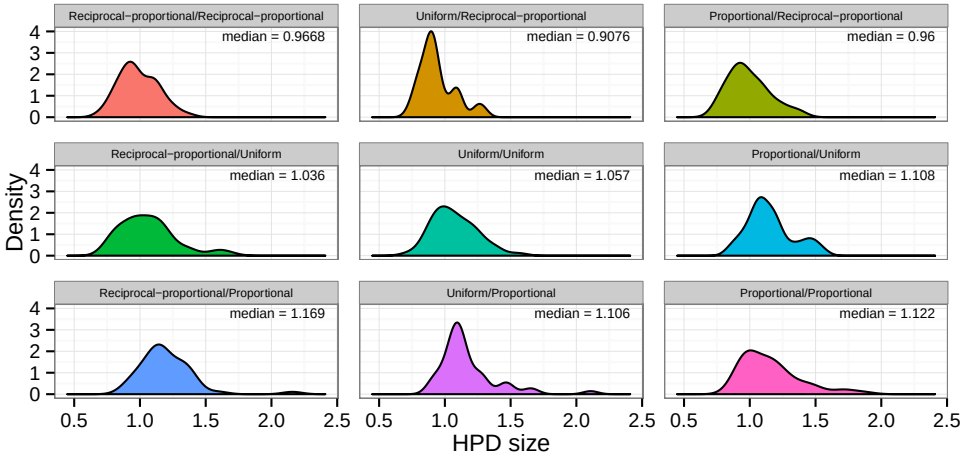
Figure 3.20: Overlaid median lines for 50 reconstructed skygrid plots for scenario 6. Each plot is for a single sampling scheme with a temporal and a spatial component, given as temporal/spatial; P=proportional, U=uniform, R=reciprocal-proportional. The red line is the true population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.21: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 6: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme, given as temporal component/spatial component.

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	0.8 (0.85)							
R/P	0.7 (0.689)	0.76 (1)						
U/R	0.76 (0.262)	0.9 (0.991)	0.74 (0.999)					
U/U	**0.5** (3.49×10^{-4})	0.68 (0.0883)	0.6 (0.177)	0.76 (0.563)				
U/P	**0.22** (2.4×10^{-12})	**0.48** (7.75×10^{-8})	**0.34** (4.12×10^{-7})	**0.5** (1.51×10^{-5})	0.56 (0.0527)			
P/R	**0.26** (7.06×10^{-11})	**0.42** (1.17×10^{-6})	**0.34** (5.49×10^{-6})	**0.44** (1.49×10^{-4})	0.62 (0.176)	0.74 (1)		
P/U	**0.2** (8.4×10^{-14})	**0.4** (7.62×10^{-11})	**0.32** (5.38×10^{-10})	**0.36** (3.86×10^{-8})	**0.58** (1.23×10^{-3})	0.82 (0.982)	0.7 (0.856)	
P/P	**0.16** (8.34×10^{-14})	**0.2** (7.66×10^{-14})	**0.24** (9.48×10^{-14})	**0.2** (4.2×10^{-13})	**0.4** (2.94×10^{-7})	0.62 (0.155)	**0.62** (0.0446)	0.76 (0.783)

(a) Percent error

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	**0.44** (2.54×10^{-3})							
R/P	**0.26** (6.87×10^{-9})	0.62 (0.273)						
U/R	0.76 (0.899)	0.62 (0.208)	**0.4** (2.36×10^{-5})					
U/U	**0.22** (5.98×10^{-12})	**0.52** (0.0178)	0.76 (0.985)	**0.34** (8.16×10^{-8})				
U/P	**0.06** (1.06×10^{-13})	**0.32** (4.11×10^{-8})	**0.54** (7.51×10^{-3})	**0.2** (7.77×10^{-14})	0.72 (0.162)			
P/R	**0.5** (1.2×10^{-4})	0.78 (0.999)	0.68 (0.73)	**0.66** (0.0323)	0.62 (0.136)	**0.52** (2.45×10^{-6})		
P/U	**0.04** (9.94×10^{-14})	**0.3** (9.27×10^{-9})	**0.56** (3.06×10^{-3})	**0.16** (6.57×10^{-14})	0.72 (0.0894)	0.84 (1)	**0.5** (6.47×10^{-7})	
P/P	**0.04** ($< 2 \times 10^{-16}$)	**0.24** (2.02×10^{-13})	**0.34** (1.25×10^{-6})	**0.12** (8.09×10^{-14})	**0.5** (2.25×10^{-4})	0.76 (0.651)	**0.34** (1.61×10^{-11})	0.76 (0.799)

(b) Percent bias

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	0.9 (0.753)							
R/P	**0.56** (1.64×10^{-6})	**0.64** (5.01×10^{-3})						
U/R	0.72 (0.775)	**0.68** (0.0234)	**0.38** (3.88×10^{-11})					
U/U	0.82 (0.337)	0.88 (1)	**0.68** (0.0422)	**0.56** (2.41×10^{-3})				
U/P	**0.52** (2.8×10^{-4})	0.56 (0.124)	0.72 (0.985)	**0.4** (3.48×10^{-8})	0.62 (0.437)			
P/R	1.02 (1)	0.86 (0.727)	**0.52** (1.3×10^{-6})	0.74 (0.798)	0.8 (0.313)	**0.5** (2.32×10^{-4})		
P/U	**0.56** (4.38×10^{-4})	0.68 (0.158)	0.84 (0.974)	**0.42** (6.41×10^{-8})	0.8 (0.505)	0.74 (1)	**0.58** (3.65×10^{-4})	
P/P	**0.72** (1.94×10^{-3})	0.78 (0.326)	0.72 (0.878)	**0.48** (5.11×10^{-7})	0.88 (0.742)	0.62 (1)	**0.72** (1.64×10^{-3})	0.74 (1)

(c) HPD size

Table 3.10: Estimated coefficients of overlapping for distributions of statistics in scenario 6. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface. Sampling schemes are given as temporal/spatial, where P=proportional, U=uniform, R=reciprocal-proportional. Figures in red compare proportional and reciprocal temporal sampling schemes for the same spatial scheme; figures in blue compare proportional and reciprocal-proportional spatial schemes for the same temporal scheme.

Scenario 7: Structured population, short period oscillations

When the default BEAST prior distributions were used on this dataset, the result was a repeat of the tendency that I first observed in the sample size analysis of scenario 4, in which the oscillations were frequently damped in the skygrid reconstructions, to the extent that, if viewed by eye, the reconstruction suggested that the EPS was invariant. While the prevalence of this phenomenon did appear to vary somewhat with sampling scheme, I considered this of little interest if the analysis could be reconfigured to avoid it. Examination of the posterior distributions for the skygrid parameters suggested that this behaviour was associated with high posterior values for the precision parameter, which governs the amount of correlation between the EPS in one interval and its neighbours [52]. For example, figure 3.22 displays 50 reconstructed graphs when reciprocal-proportional temporal sampling and uniform spatial sampling were used. The graphs are annotated with the posterior median value of the precision parameter. It is clear that sampling replicates for which the graph is damped are those for which the estimate is large. As this suggested a misspecified prior distribution, all datasets were reanalysed with the default Gamma$(0.001, 0.001)$ prior on the precision parameter replaced by a more informative Gamma$(0.1, 0.1)$ distribution.

When analyses were rerun with this modification, the overwhelming majority of reconstructed plots displayed clear oscillatory dynamics. Figure 3.23 displays the overlaid median lines, and figure 3.24 the KDEs. The figures for overlapping coefficients suggest a similar picture to scenario 6 (table 3.11); moving from a proportional to reciprocal-proportional temporal scheme reduces both error and bias, whereas doing the same for spatial schemes reduces bias but there is little evidence that it does so for error. There is, however, evidence for a reduction of HPD size in both cases, whereas in scenario 6 this was only true when the temporal scheme was varied. Unlike scenario 4, in which the dynamics of the total

population obeyed the same curve as this scenario, the best-performing sampling schemes were reciprocal-proportional, not uniform.

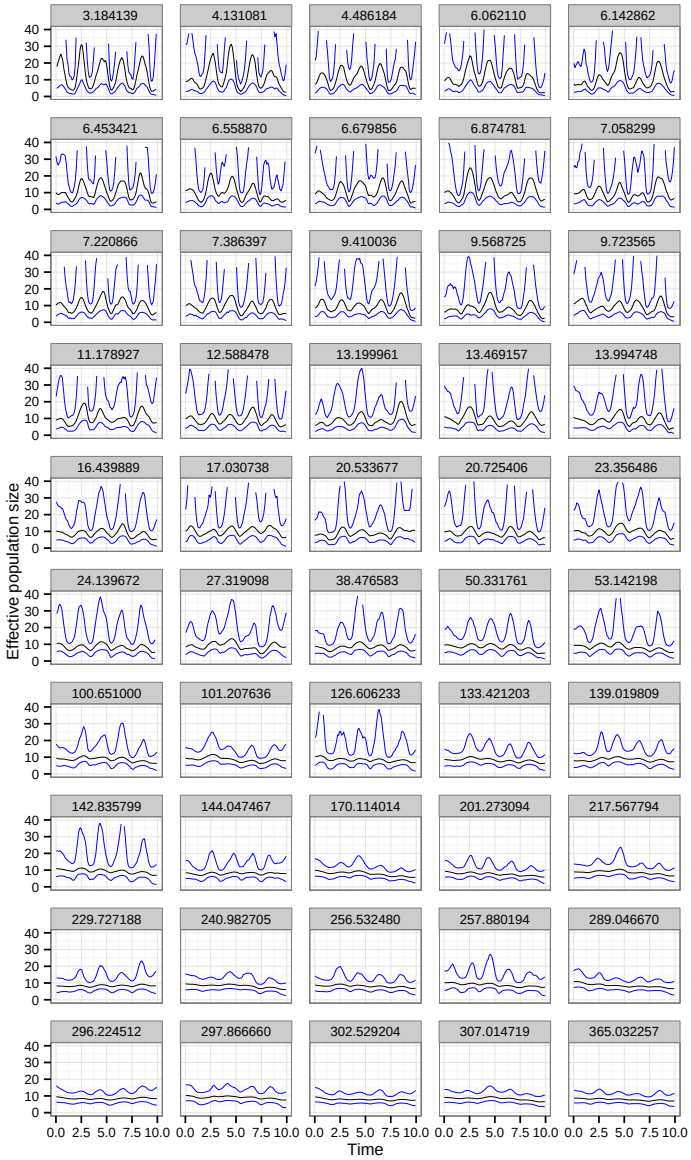
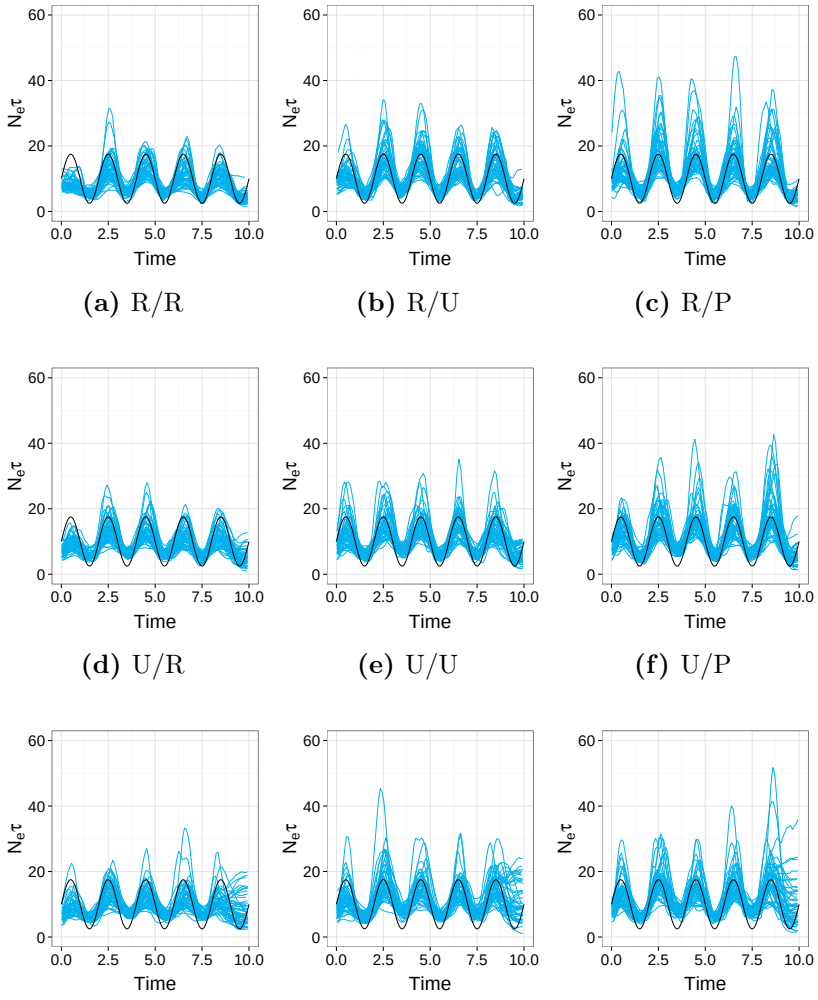


Figure 3.22: Skygrid reconstructions for an initial analysis of 50 replicates of the reciprocal-proportional/uniform scheme in scenario 7, labelled and sorted by posterior median value of the skygrid precision parameter. The black line is the median estimate, the blue lines the bounds of the 95% HPD interval.

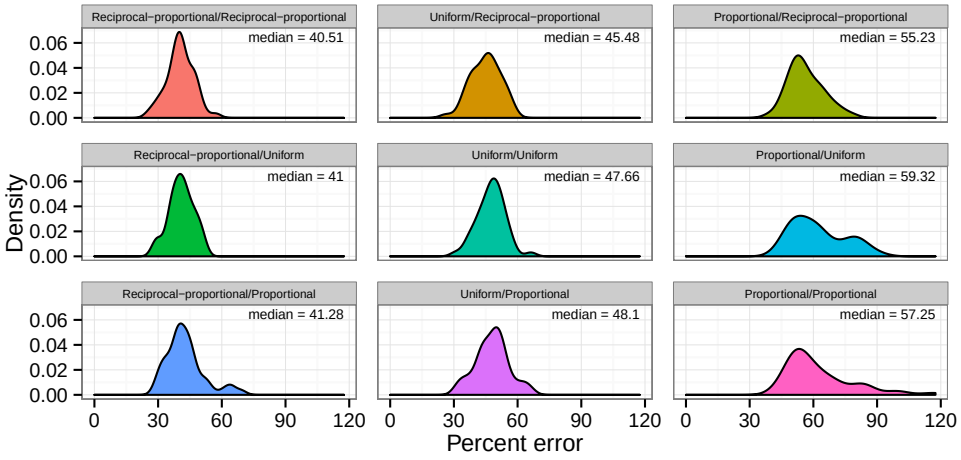


(a) R/R **(b)** R/U **(c)** R/P

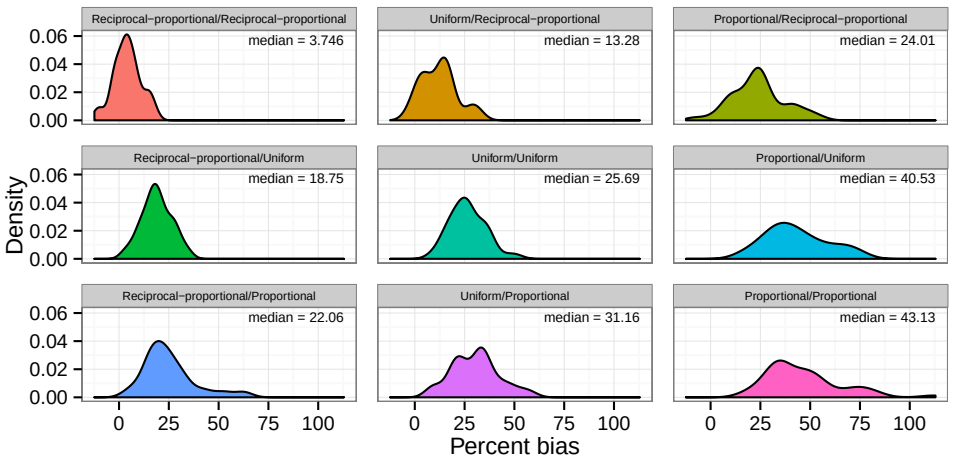
(d) U/R **(e)** U/U **(f)** U/P

(g) P/R **(h)** P/U **(i)** P/P

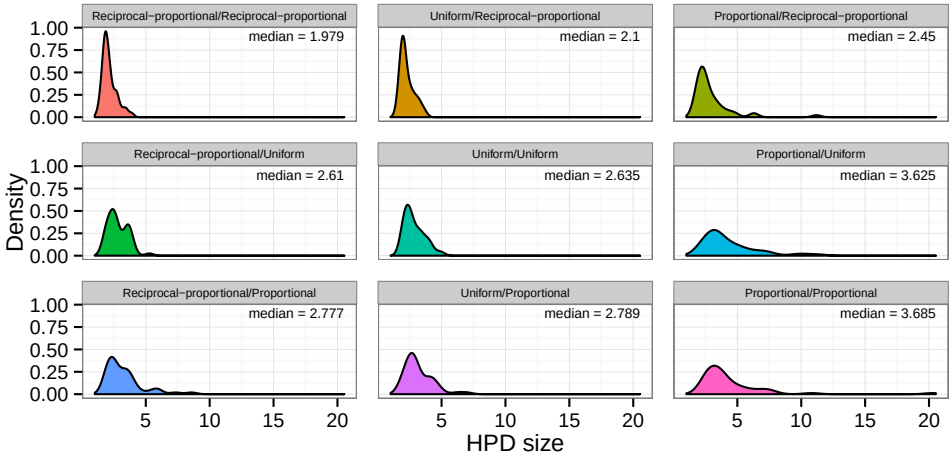
Figure 3.23: Overlaid median lines for 50 reconstructed skygrid plots for scenario 7. Each plot is for a single sampling scheme with a temporal and a spatial component, given as temporal/spatial; P=proportional, U=uniform, R=reciprocal-proportional. The red line is the true population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.24: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 7: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme, given as temporal component/spatial component.

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	0.86 (1)							
R/P	0.8 (0.983)	0.84 (0.995)						
U/R	0.68 (0.213)	0.72 (0.304)	0.7 (0.852)					
U/U	**0.56** (2.24×10^{-3})	**0.58** (4.49×10^{-3})	0.6 (0.0774)	0.86 (0.882)				
U/P	**0.62** (9.62×10^{-4})	**0.6** (2.01×10^{-3})	**0.64** (0.043)	0.82 (0.781)	0.88 (1)			
P/R	**0.24** (1.25×10^{-13})	**0.26** (1.9×10^{-13})	**0.32** (4.61×10^{-11})	**0.48** (8.22×10^{-7})	**0.56** (1.12×10^{-3})	**0.62** (2.59×10^{-3})		
P/U	**0.18** (9.86×10^{-14})	**0.2** (1.18×10^{-13})	**0.24** (1.11×10^{-13})	**0.44** (7.09×10^{-10})	**0.46** (4.93×10^{-6})	**0.5** (1.45×10^{-5})	0.78 (0.971)	
P/P	**0.18** (1.19×10^{-13})	**0.18** (6.63×10^{-14})	**0.3** (1.56×10^{-13})	**0.38** (2.65×10^{-9})	**0.56** (1.39×10^{-5})	**0.56** (3.93×10^{-5})	0.82 (0.992)	0.9 (1)

(a) Percent error

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	**0.3** (5.02×10^{-5})							
R/P	**0.22** (7.01×10^{-10})	0.82 (0.621)						
U/R	0.56 (0.313)	0.6 (0.246)	**0.48** (4.4×10^{-4})					
U/U	**0.2** (5.54×10^{-13})	0.64 (0.0943)	0.8 (0.987)	**0.38** (3.16×10^{-6})				
U/P	**0.1** (6.83×10^{-14})	**0.56** (2.01×10^{-3})	0.7 (0.461)	**0.3** (3.32×10^{-9})	0.78 (0.97)			
P/R	**0.38** (1.68×10^{-9})	0.7 (0.708)	0.76 (1)	**0.44** (7.75×10^{-4})	0.72 (0.972)	0.7 (0.376)		
P/U	**0.04** ($< 2 \times 10^{-16}$)	**0.24** (1.41×10^{-10})	**0.48** (1.55×10^{-5})	**0.18** (8.32×10^{-14})	**0.54** (1.59×10^{-3})	0.66 (0.0809)	**0.42** (7.89×10^{-6})	
P/P	**0.04** ($< 2 \times 10^{-16}$)	**0.18** (6.59×10^{-12})	**0.38** (1.58×10^{-6})	**0.16** (8.98×10^{-14})	**0.54** (2.5×10^{-4})	**0.64** (0.0226)	**0.36** (7.63×10^{-7})	0.86 (1)

(b) Percent bias

	R/R	R/U	R/P	U/R	U/U	U/P	P/R	P/U
R/U	**0.52** (6.8×10^{-4})							
R/P	**0.54** (3.28×10^{-6})	0.86 (0.977)						
U/R	0.82 (0.983)	**0.64** (0.0331)	**0.64** (5.4×10^{-4})					
U/U	**0.54** (2.21×10^{-4})	0.76 (1)	0.8 (0.996)	**0.64** (0.0145)				
U/P	**0.44** (2.41×10^{-7})	0.6 (0.839)	0.66 (1)	**0.54** (6.4×10^{-5})	0.8 (0.936)			
P/R	**0.62** (8.95×10^{-3})	0.76 (0.999)	0.78 (0.734)	0.66 (0.189)	0.82 (0.993)	0.64 (0.421)		
P/U	**0.3** (9.69×10^{-14})	**0.52** (1.39×10^{-3})	0.64 (0.064)	**0.36** (8.18×10^{-12})	**0.66** (3.81×10^{-3})	0.66 (0.203)	**0.56** (7.53×10^{-5})	
P/P	**0.3** (8.07×10^{-14})	**0.56** (5.03×10^{-4})	**0.68** (0.0314)	**0.36** (1.42×10^{-12})	**0.66** (1.47×10^{-3})	0.64 (0.116)	**0.56** (2.32×10^{-5})	0.86 (1)

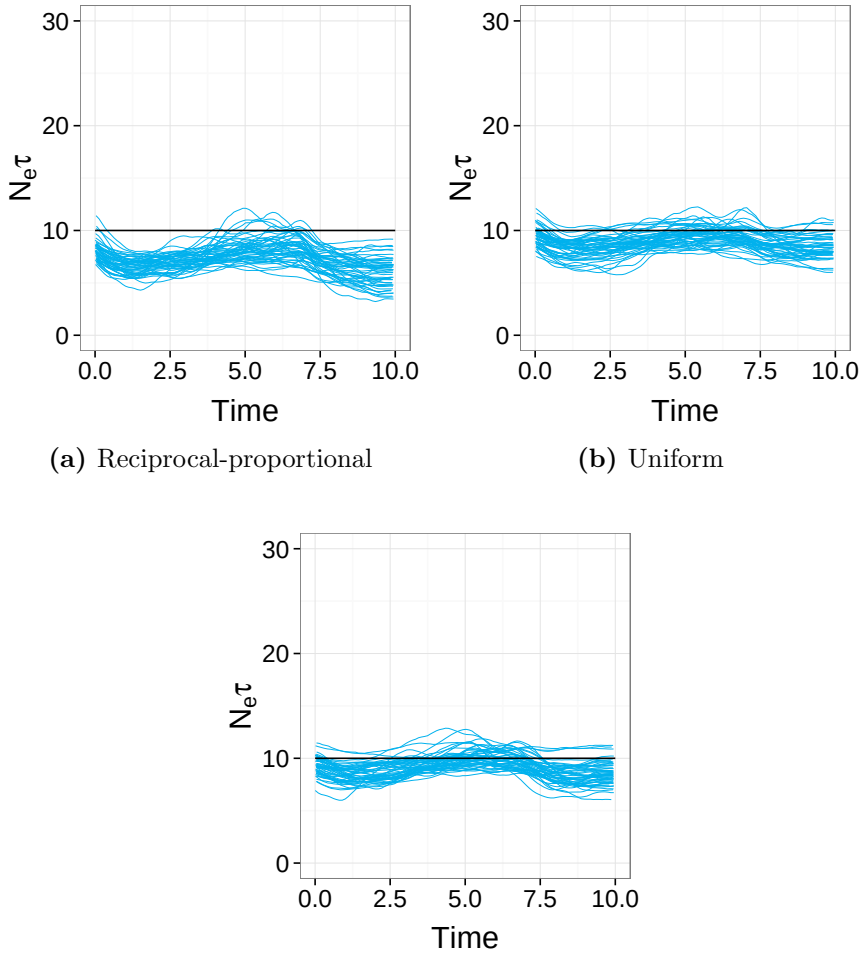
(c) HPD size

Table 3.11: Estimated coefficients of overlapping for distributions of statistics in scenario 7. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface. Sampling schemes are given as temporal/spatial, where P=proportional, U=uniform, R=reciprocal-proportional. Figures in red compare proportional and reciprocal temporal sampling schemes for the same spatial scheme; figures in blue compare proportional and reciprocal-proportional spatial schemes for the same temporal scheme.

Scenario 8: Structured population, short period oscillations, out of phase

The total population size being constant through time in this scenario, I varied only the spatial sampling scheme. In contrast to any other scenario examined here, the bias is towards underestimating sizes (figure 3.25, figure 3.26). Notably the bias is most serious for the reciprocal-proportional scheme.

The final set of coefficients of overlapping and post-hoc test p-values are given in

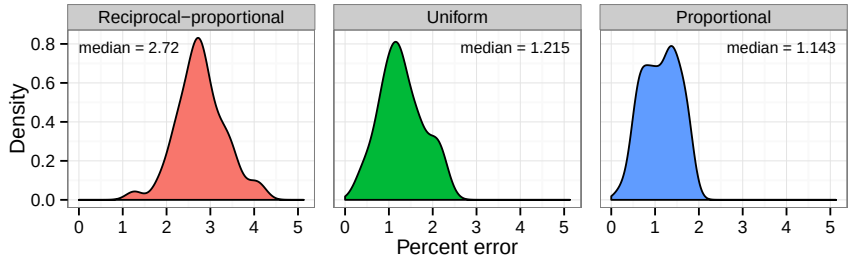


(a) Reciprocal-proportional

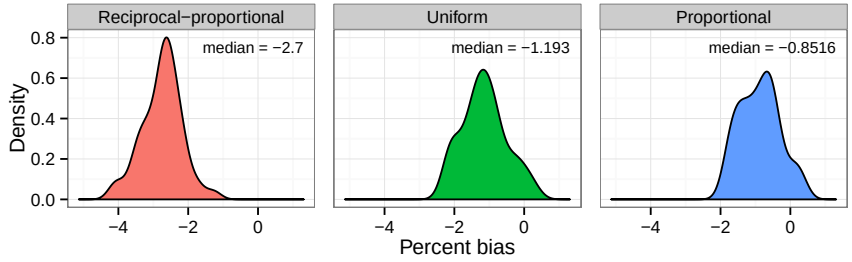
(b) Uniform

(c) Proportional

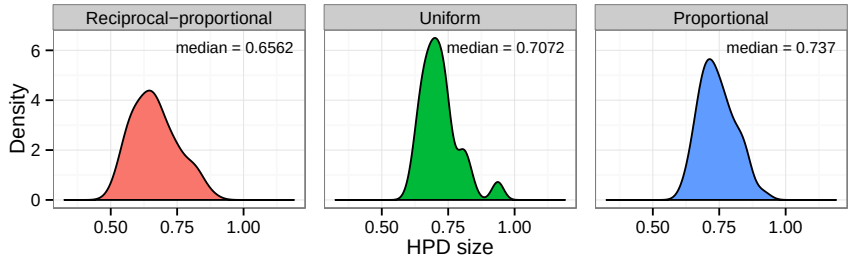
Figure 3.25: Overlaid median lines for 50 reconstructed skygrid plots for scenario 8. The red line is the true population size.



(a) Percent error



(b) Percent bias



(c) HPD size

Figure 3.26: Kernel density estimates for the distribution of statistics indicating the accuracy and precision of the skygrid reconstructions in scenario 8: a) percent error, b) percent bias, c) HPD size. Each plot corresponds to and is labelled with a different sampling scheme.

table 3.12. There is very little to separate the uniform and proportional schemes, but the poor performance of reciprocal-proportional sampling is evident. The latter does, however, still show the superior precision found in other scenarios.

	Reciprocal-proportional	Uniform
Uniform	**0.14** (5.04×10^{-14})	
Proportional	**0.06** (2.23×10^{-14})	0.8 (0.607)

(a) Percent error

	Reciprocal-proportional	Uniform
Uniform	**0.18** (7.28×10^{-14})	
Proportional	**0.06** (3.73×10^{-14})	0.82 (0.31)

(b) Percent bias

	Reciprocal-proportional	Uniform
Uniform	**0.64** (6.62×10^{-3})	
Proportional	**0.52** (2.51×10^{-6})	0.8 (0.144)

(c) HPD size

Table 3.12: Estimated coefficients of overlapping for distributions of statistics in scenario 8. Each entry in each table compares a statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.

3.3.2 Phylogeographical reconstruction

Scenario 5: Structured population, constant size

The first point of note is that transition rate estimates were extremely noisy. As a demonstration of this, I constructed a KDE of the posterior distribution of

each between-deme rate for each replicate of each sampling scheme, and then calculated the coefficient of overlapping of the KDEs from each pair of replicates of the same scheme. This was done for the analysis without BSSVS only: since posterior distributions of rates from BSSVS usually have many repeated values of 0, kernel density estimation is not appropriate. The results of this can be seen in figure 3.27; each histogram represents the distribution of coefficients over all pairs of replicates, and is annotated with the quartiles of this distribution. For most rates the median coefficient was in the range 0.7-0.8, and values below 0.5 were not uncommon; when comparing all rate estimates from all replicates, the proportion of pairwise comparisons for which it was less than this was 0.1 for proportional sampling, 0.098 for uniform, and 0.0903 for reciprocal-proportional. The very lowest coefficient of all was 0.0505. Clearly, the stochastic choice of samples for inclusion has a major effect on the estimates of these parameters.

Histograms for Kendall's τ statistic, measuring correlation between MAP estimates for rates from the analyses and the true values of each rate (the M_{ij} described under Methods), can be seen in figure 3.28. While there is evidence to suggest that the distribution of estimates from proportional sampling is different from that of the other two schemes (table 3.15), there is considerable overlap between the histograms and any such effect appears small. If the τ statistic was used as the basis for a hypothesis test for the existence of a correlation, the resulting p-value would be less than 0.05 in all but five replicates (two for proportional sampling, one for uniform and two for reciprocal-proportional).

I investigated the effect of sample size on rate estimates by drawing 10 replicates each of the uniform sampling scheme for sample sizes ranging from 25 to 500 in increments of 25, and, as before, used least-squares regression to fit a model of the relationship between τ and sample size. AICc values are given in table 3.14. The logarithmic model was preferred with an exponential error model (although there was practically no difference in AICc value between this and the power law

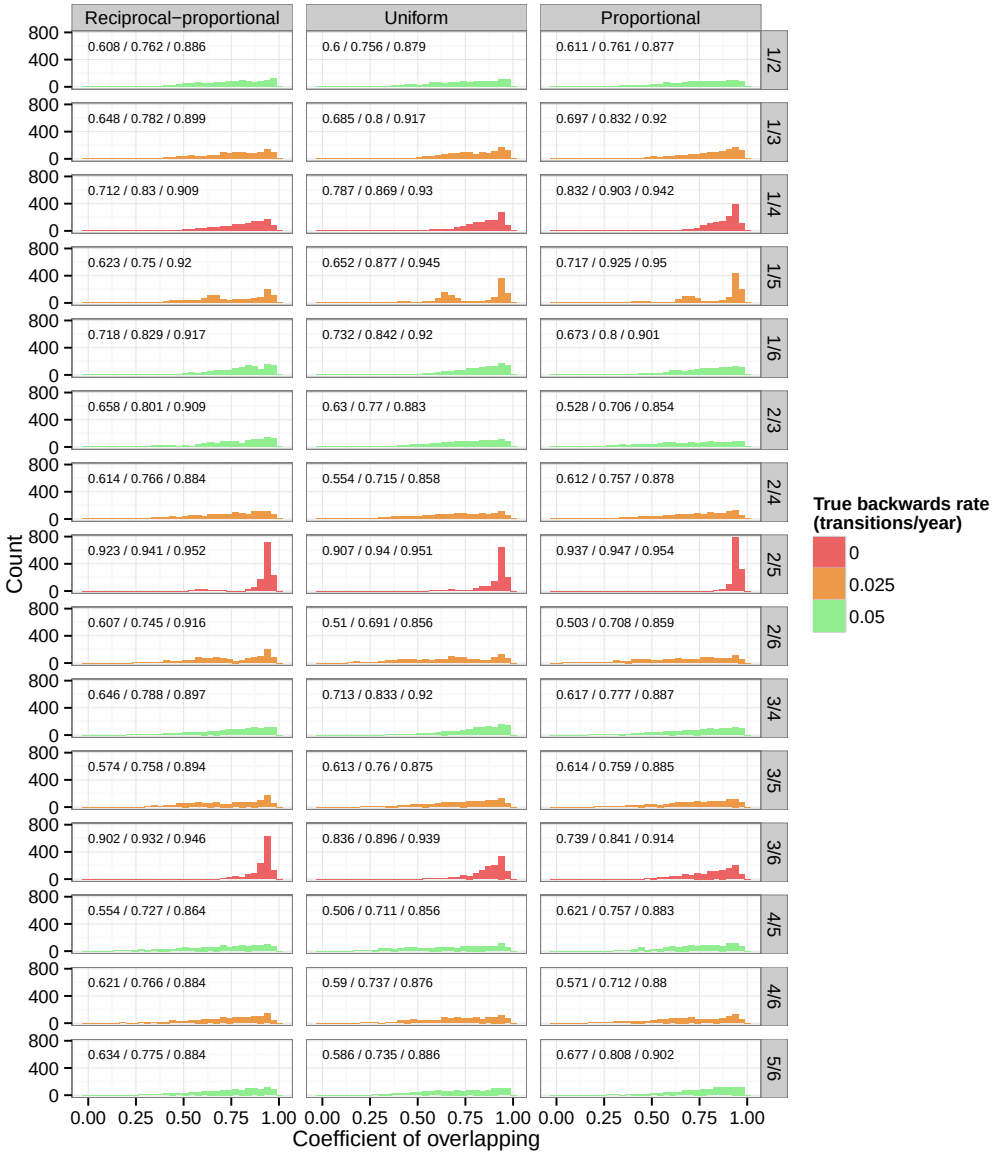


Figure 3.27: Illustration of the noisiness of deme-to-deme transition rate estimates. Each histogram corresponds to a single between-deme rate (a rate of transition between two demes in figure 3.1) and a single sampling scheme. The data is the set of coefficients of overlapping for KDEs estimating the posterior distribution of that rate, for every pair of replicates of that sampling scheme (as there were 50 replicates per scheme, there are 1225 pairs). Histograms are coloured by the true rate; labels on the right refer to the two numbered demes that the rate is between. Each is labelled with 25th, 50th and 75th percentile values of the coefficient.

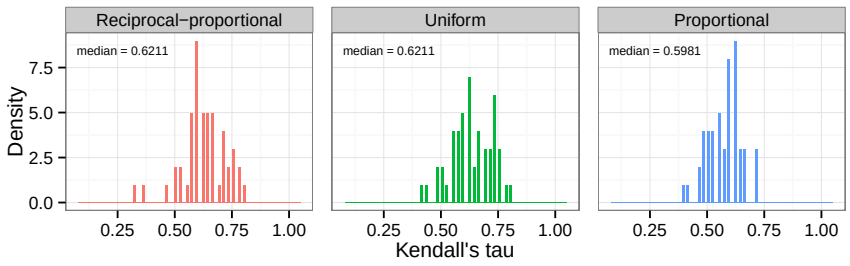
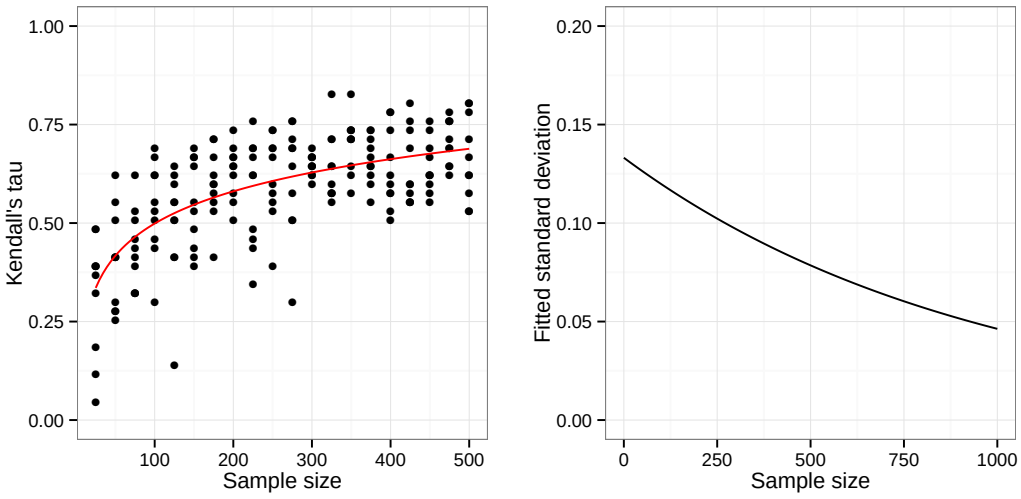


Figure 3.28: Histograms for Kendall's τ statistic, for correlation between point estimates of between-deme rates and the true rates, in the phylogeography analysis of scenario 5.

	Reciprocal-proportional	Uniform
Uniform	0.5 (0.988)	
Proportional	**0.62** (9.53×10^{-3})	**0.66** (5.85×10^{-3})

Table 3.13: Estimated histogram intersection statistics for distributions of Kendall's τ statistic for the correlation between maximum a posteriori probability estimates for between-deme transition rates, and the actual rates used to generate the simulation. Each entry compares the statistic between two sampling schemes. Numbers in parentheses are p-values from post-hoc (Nemenyi) tests for the null hypothesis that the data used to estimate each KDE came from the same distribution; where these are < 0.05 the coefficient of overlapping is given in boldface.

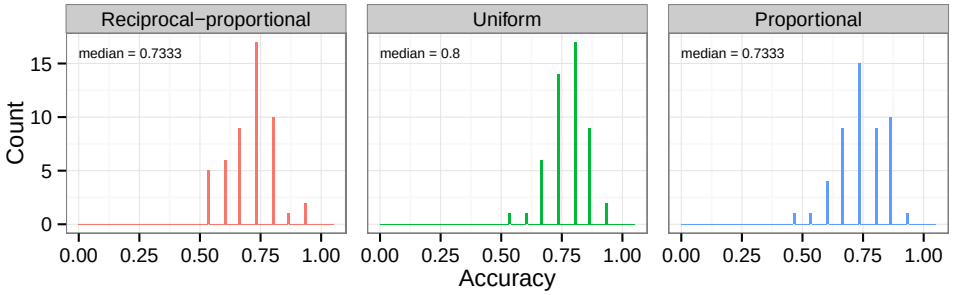


(a) **(b)**

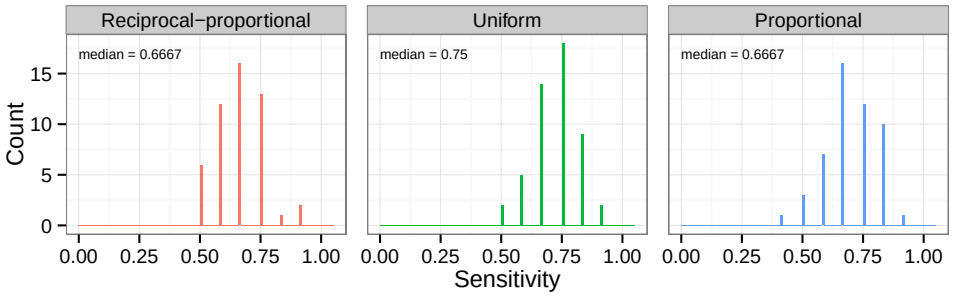
Figure 3.29: a) Scatter plot of Kendall's τ statistic (for the correlation between maximum a posteriori probability estimates for between-deme transition rates, and the actual rates used to generate the simulation) and sample size for 100 replicates of the uniform sampling scene in scenario 4. The red line represents the best-fit model determined by weighted least squares regression and corrected Akaike information criterion. b) Curve of the standard deviation function of the best-fit model.

error model); the best-fit parameters were $A = 0.117$, $B = -0.044$, $\sigma^2 = 0.0178$ and $t = -1.06 \times 10^{-3}$. It can be seen that, in contrast to previous sample size exercises, considerable gains in τ, and more consistent estimates, would continue to be found as the sample size was increased above 500 if this relationship still held (figure 3.29).

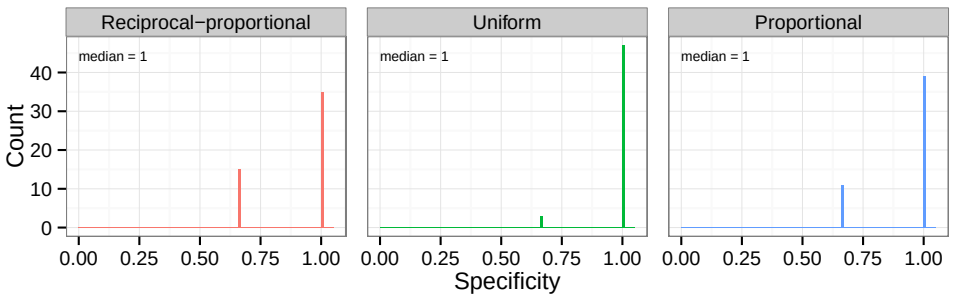
Turning to the use of BSSVS as a method of determining which between-deme rates are zero, I tested the performance of the three sampling schemes by calculating, for each replicate, the overall accuracy, sensitivity and specificity of using BF>3 as a binary classifier for the presence of a nonzero rate. The results are summarised as histograms in figure 3.30.



(a) Accuracy



(b) Sensitivity



(c) Specificity

Figure 3.30: Histograms of the overall accuracy, sensitivity, and specificity of the use of Bayes Factor > 3 in a BSSVS reconstruction to identify nonzero rates of movement between demes, by sampling scheme, phylogeography analysis of scenario 5.

Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
$\tau = An + B$	-277.81	-290.95	-295.00	-293.12
$\tau = A\ln(n) + B$	-320.80	-328.16	-328.09	-325.98
$\tau = \frac{A}{n} + B$	-308.82	-321.21	-322.85	-320.75
$\ln(\tau) = An + B$	-135.46	-237.89	-260.27	-258.42
$\ln(\tau) = A\ln(n) + B$	-183.13	-271.59	-285.95	-283.85

Table 3.14: AICc values for models of the relationship between Kendall's τ statistic (for the correlation between maximum a posteriori probability estimates for between-deme transition rates, and the actual rates used to generate the simulation), and sample size (n), scenario 5, whose parameters were fit by least-squares regression.

Figures for histogram intersection and the results of post-hoc tests can be be found in table 3.15. There is considerable intersection, but the results do give evidence of a difference in performance between the uniform and reciprocal-proportional schemes for both overall accuracy and sensitivity, which would be in favour of the former.

All observations from the same sampling scheme were then pooled, and the resulting sets used to construct a receiver-operator curve (ROC) using BF values as cut-offs. This can be seem in figure 3.31. BSSVS performs best as a classifier when the sampling scheme is uniform. Two particular points are marked on each curve. One corresponds to BF=3, the most widely-used cutoff for analyses of this sort. As can be seen, using this value favours specificity at the expense of sensitivity, at least for this simulated population. The other corresponds to the BF value that maximises accuracy. These are all less than 1; an analysis using these values will support all links with posterior odds higher than prior odds, and some links with posterior odds *lower* than prior odds.

	Reciprocal-proportional	Uniform
Uniform	**0.66** (4.93×10^{-3})	
Proportional	0.8 (0.213)	0.82 (0.325)

(a) Accuracy

	Reciprocal-proportional	Uniform
Uniform	**0.7** (0.0163)	
Proportional	0.8 (0.183) 0.0163	0.86 (0.590)

(b) Sensitivity

	Reciprocal-proportional	Uniform
Uniform	0.76 (0.366)	
Proportional	0.92 (0.894)	0.84 (0.640)

(c) Specificity

Table 3.15: Histogram intersection values and results of the post-hoc tests for the performance of BSSVS as a binary classifier in scenario 5. Figures, which have been adjusted for multiple testing, are p-values for evidence against the null hypothesis that the distribution of a statistic is the same across two sampling schemes.

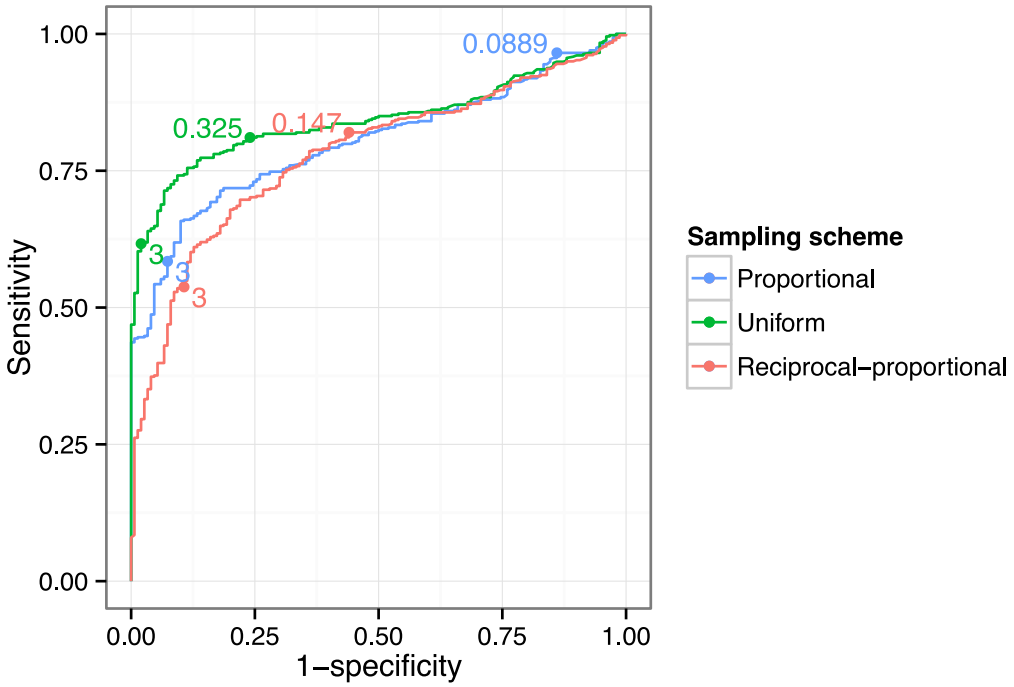


Figure 3.31: ROC curves for the performance of BSSVS as a classifier for zero or nonzero rates. Each line corresponds to a sampling scheme. The two points marked on each correspond to, and are marked with, the Bayes factor that maximises the accuracy, and a Bayes factor of 3.

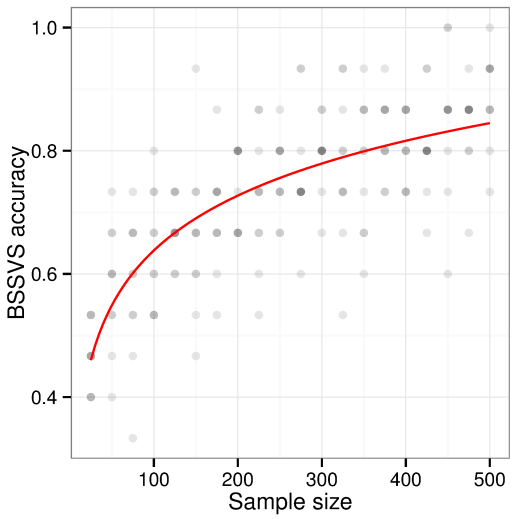


Figure 3.32: Scatter plot of the accuracy of BSSVS in identifying zero and nonzero rates of movement against sample size. Many points overlap, and darker points indicate the presence of more than one replicate with this accuracy. The red curve is the best model of the relationship between sample size and accuracy, fit by least-squares regression.

Figure 3.32 plots the accuracy of the BSSVS classification (with BF=3) against sample size. The red curve was again fit by least-squares regression; AICc values for various models are in table 3.16. The best model is given by a logarithmic curve $g(n) = 0.128\log(n) + 0.0466$. The plot does not suggest heteroscedasticity and, indeed, the best variance model does not include it, so the fitted variance is constant at $\sigma^2 = 7.72 \times 10^{-3}$.

When ROC curves are drawn for each sample size, again pooling the results of all replicates, the performance as a classifier shows an obvious improvement as sample size increases (figure 3.33). The area under the curve (AUC) statistic, indicative of the performance of the classifier, strongly suggests a linear relationship with sample size over the range of sizes tested (figure 3.34, $r^2 = 0.944$ for simple linear regression).

Model	$v(n)$			
	1	e^{tn}	$\|n\|^t$	$t_1 + \|n\|^{t_2}$
error $= An + B$	-354.53	-354.74	-355.40	-353.48
error $= A\ln(n) + B$	-384.89	-383.01	-382.81	-380.71
error $= \frac{A}{n} + B$	-342.46	-340.90	-341.27	-339.17
$\ln$(error) $= An + B$	-318.71	-337.29	-341.55	-339.64
$\ln$(error) $= A\ln(n) + B$	-364.63	-373.50	-372.34	-370.24

Table 3.16: AICc values for models of the relationship between accuracy of BSSVS as a classifier of zero and nonzero rates at BF=3 and sample size, whose parameters were fit by least-squares regression.

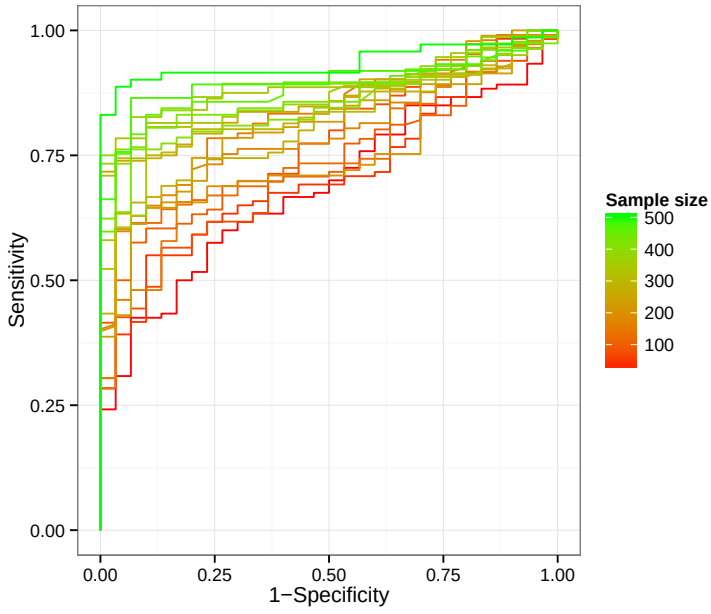


Figure 3.33: ROC curves for the performance of BSSVS as a classifier for zero or nonzero rates. Each line is for observations from analyses of a different sample size.

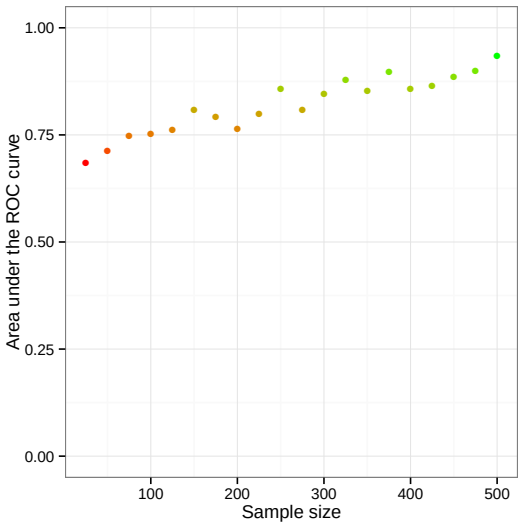


Figure 3.34: The area under the ROC curves in figure 3.33 as a function of sample size.

3.4 Discussion

The simulation exercise detailed in this chapter is a more comprehensive effort than any previously published to investigate the effects that sampling schemes have on the reconstruction of spatial and temporal dynamics from nucleotide sequences. Caution must be taken in generalising the results here. The range of demographic scenarios that could be simulated is effectively limitless, and attempting to cover every possible complication or nuance is not feasible. I also assumed a simple, and invariant, mutation model. Since discrete-traits phylogeography commonly treats large geographical units such as countries as traits, it is also a great simplification to model this under a structured coalescent by assuming that all lineages within a location mix freely; this assumption is even worse if the trait is something other than a geographical entity, such as a host species. Researchers wishing to investigate sampling effects in a situation analogous to a particular study that they are conducting may wish to design similar simulations with population structures

that are more appropriate for their work. Nevertheless, there are several results of this analysis which should inform sampling strategies in general.

I made the choice to analyse multiple replicates of sampling schemes drawing from the same pool of sequences, rather than taking the approach of previous work [23, 64] in which separate coalescent simulations were performed from the same collection of tips; in that situation every sampling "replicate" actually has a unique phylogenetic tree. I felt my approach was more reflective of the process of devising an actual scheme in the real world, and added an element of stochastic variation in the process of sample collection. A drawback is that each master set is a unique stochastic realisation of the coalescent simulation, and as a result, some features may be unique only to that realisation. This is presumably why, for example, there is consistently more variation in the estimation of some transition rates than others even where their actual values are identical (figure 3.27). Nevertheless, all 8 scenarios considered here used a different master set, and many of the phenomena noted are consistent across them.

It is certainly concerning that stochastic variation in the sequences picked by a sensible sampling scheme can nonetheless introduce spurious temporal variation in skygrid reconstructions, to the extent that hypothesis tests can actually reject an accurate simple model in favour of a more complicated one, although the latter phenomenon was basically absent in three of the five examples tested and only overwhelming in one. I make two recommendations as a result of this. The first is that the behaviour of the median line in a plot from the Bayesian skyline family should be regarded with scepticism, and certainly the HPD region must be taken into account. For example, the median lines of the bottom five graphs in figure 3.5 could lead an unwary researcher to suggest many potentially interesting historical scenarios, all of which would in fact be entirely sampling artefacts. But in almost every case amongst the 50 in that figure, a straight line representing an invariant EPS could be drawn through the HPD bounds over the entire extent of the graph.

On the other hand, most graphs in figure 3.14 which display no damping (although not all) could not accept such a line. The second recommendation, especially when using a random procedure to downsample large datasets of sequences collected in the past, is to compare the results of analyses from more than one replicate of the scheme, in case any distinctive features of the reconstruction are no more than the results of the samples chosen.

One practice that certainly should not continue is the presentation of demographic reconstructions based on datasets consisting disproportionately of a collection of recent isolates taken from the same area. This clearly introduces a spurious bottleneck effect. While this observation is not new [23, 64], it has only been previously shown in situations where all tips in the tree are contemporaneous, and those papers have rarely been cited in the pathogen phylogenetics literature; here I have confirmed that it does also hold when tips are distributed over a wider time period.

The apparent superiority of reciprocal-proportional schemes for skygrid reconstruction in many scenarios, which would suggest that the best strategy would be to include epidemiologically rare isolates at a greater frequency than more common ones, is highly unexpected. Nevertheless, I recommend caution in adopting it (leaving aside, for now, the practical considerations involved in calculating the reciprocal of the EPS), firstly because there was no suggestion that the rule held for discrete traits analysis, and secondly because it was not superior in every scenario. For exponential growth (scenario 2) the reason for its poor performance is presumably that which was identified by de Silva et al. [31]; the plot will tail off when there are few samples left in the dataset, and the more quickly the remaining number declines, the earlier this will happen. This is an unfortunate feature of the reconstruction of the dynamics of epidemic using such methods, as one would like to be able to detect when the growth phase has peaked, but a spurious tailing off is to be expected. For complicated dynamics (scenario 8), however, the reason

is less clear, and since real-world scenarios will be yet more complex than any considered here, it may be premature to attempt to use a scheme of this sort.

Identifying the reasons for superior performance of the reciprocal-proportional scheme when it does occur would likely require analytical work that is beyond the scope of this chapter. One possibility does present itself: it is clear from scenario 5 that oversampling small populations decreases EPS estimates. In almost all these scenarios the EPS tended to be overestimated regardless of sampling scheme; it may be that all the reciprocal-proportional scheme, which by definition oversamples small populations, does is mitigate this bias. If we compare the results for bias in scenario 3 (which has no spatial component) with those in scenario 6 for reciprocal-proportional spatial sampling, the latter are actually less biased even though the population model is inappropriate; the same goes when comparing scenario 4 to scenario 7. In scenario 8, on the other hand, where the overall bias is in the opposite direction, reciprocal-proportional sampling just makes matters worse.

The bias towards overestimating EPSs, even in scenario 1, is itself unexpected, as the original skyline has been shown to be unbiased as an estimator of the harmonic mean of the EPS on an interval [116], and of course if the population size is constant then the harmonic mean size is just the value of this size. While they did not formally investigate it, such a bias was not reported in the methodological papers introducing the skygrid [52] or indeed the skyride [102], and while Heller et al. [64] did report it as the result of population structure that the model did not take account of, I found it even when that was not present. It transpires that this is the result of the process of simulating large trees under a particular demographic model and then taking a small subsample of the tips for analysis; figure 3.35b displays overlaid plots for another 50 replicates from scenario 1, except that this time each tree was simulated individually from 300 tips. The median percent bias in this case is only -0.0524, much smaller than the 0.982 from uniform sampling in

scenario 1. Figure 3.35 also shows that the inaccurate behaviour of the median line in the period prior to sampling appears to also be the result of subsampling a large set of simulated sequences; in every sampling scheme of every scenario in the main text, the median line diverged upwards from the true dynamics going backwards from the first-sampled tip, but this is not true if trees are simulated individually. The reasons for these phenomena are unclear. As neither situation is very analogous to real life, the implications of this for the analysis of real data is also nebulous, but the situation in the main body of the text is somewhat more realistic, and this is of some concern. Skyline-family plots are frequently used in inference of the population dynamics of eukaryotes in the very distant past (e.g. [69, 100, 143]), and the possibility of a hitherto unknown sampling effect warrants investigation. The behaviour of the median line in the earlier period is of greater concern than an overall bias in $N_e\tau$ estimates as the latter are rarely of great interest in phylodynamics. It should be emphasised that while the median line universally displayed this behaviour when a master set was subsampled, the true dynamics rarely conflicted with the HPD region, and hence this is another reason that the behaviour of the median line should not be viewed in isolation. It is also clear that, in pathogen studies, the section of a skyline-family reconstruction that comes from the period while sampling is ongoing is most useful.

The relationship between the EPS parameter that is estimated by coalescent-based methods and parameters of epidemiological significance for infectious disease outbreaks is, as alluded to in the introduction to this chapter, not straightforward. The exact numerical estimates of $N_e\tau$ coming from coalescent analysis escape easy interpretation, given the simplistic and often-violated assumptions that are the basis for their inference, and the complicated relationships with disease dynamics [48, 153, 155]. This has consequences for attempts to select a sampling scheme based on temporal trends in disease occurrence. In this chapter I effectively assumed that this problem had been solved, and that it was possible to choose a

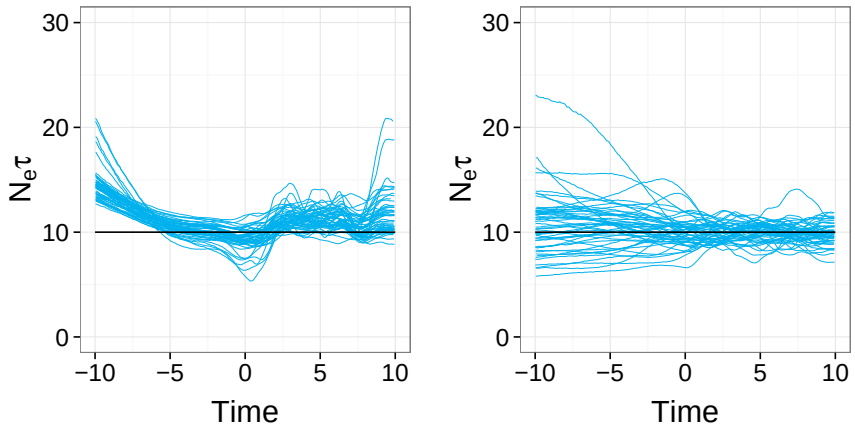


Figure 3.35: Effect on skygrid reconstruction of downsampling a large sequence dataset versus simulating a tree on a random set of tips. Subfigure a) is the same as figure 3.3a, except that the reconstruction prior to time 0 is shown. Subfigure b) comes from sequences simulated under scenario 1, but no downsampling was performed; instead, for each of the 50 replicates, a distinct phylogeny was simulated on a set of 300 tips uniformly distributed on the interval [0,10]

dataset by weighting the probability of inclusion of sequences according to some relationship between $N_e\tau$ and known temporal trends. This is naturally highly problematic in practice, but what the results here show is that it may in fact be unnecessary. The uniform sampling methods (in more standard epidemiological terminology, stratifying by time period or location) performed as well or better than proportional sampling in the vast majority of scenarios, for both skygrid and phylogeographical reconstruction. Since the occasionally superior performance of reciprocal-proportional sampling appears to be unreliable, and does not apply to phylogeography, I would recommend the uniform scheme in practice. I would caution, however, that a uniform sampling scheme must be carefully designed lest it become effectively proportional. This would occur, for example, if one stratified by year for a pathogen causing disease with a strong seasonal aspect; a random selection from a year's worth of influenza samples will probably result in a set in which most come from the winter. Care must therefore be taken not to select too wide a time window.

The papers that introduced both the skygrid [52] and its predecessor [102] noted the difficulty involved in selecting an appropriate prior distribution for the precision parameter of the Gaussian mean random field (GMRF) that dictates the relationship between the EPSs in successive time intervals. The assumption is that the logarithm of the population size in an interval is normally distributed with the mean being the log size in the previous interval, and precision ρ; the prior that I had to adjust in scenario 7 is on ρ. Since the meaning of the numerical estimates of EPS from these methods is unclear, there are few intuitions to use in selecting a suitable informed prior. Nevertheless, it is clear from this exercise that this is not merely a theoretical concern. The diffuse Gamma$(0.001, 0.001)$ prior used by default in BEAST can, by allowing the precision parameter to take very large values, effectively smooth genuine variation into a straight line. This problem occurs with greater frequency at low sample sizes, and is one reason to increase the

size of the dataset. Once again, a basic recommendation that I can make is to at the very least re-run an analysis using a different prior, particularly if the reconstructed plot appears to be flat. Further work on the proper interpretation of skyline-family EPS values may allow researchers to identify informative distributions for this parameter.

By investigating the effect of sample size, I wanted to establish the circumstances under which it would be prudent to add more sequences to the analysis. As before, caution must be taken in inferring general rules based on the specific scenarios presented here. Nevertheless, skygrid reconstructions for sample sizes of 100 or less tended to be error-prone and unreliable. For two replicates, the HPD region was very wide indeed in parts of the timeline, suggesting that the MCMC was simply unable to estimate the EPS during times with any degree of precision. When trying to reconstruct potentially complicated dynamics, such as in scenario 4, it seems worthwhile to use at least 200 sequences if not more. On the other hand, increasing from 400 to 500 sequences may not be worth the additional sequencing and computational time. While caution should always be applied in assuming that the relationships modelled by fitted curves continue beyond the bounds for the explanatory variable that were used in the fitting, in this case there is no obvious reason why behaviour should rapidly diverge for sequence counts above 500, and certainly one would not expect trends to reverse. With that in mind, there was little suggestion that moving to, for example, 600 sequences would result in substantial gains of either accuracy (as determined by percent error) or precision (as determined by HPD size). The fitted standard deviation functions suggest modest gains in reducing between-replicate variation in error by going to 500 sequences and potentially beyond (figures 3.6b and 3.15b) and virtually nothing when it comes to variation in precision beyond about 250 sequences (figures 3.7b and 3.16b).

CTMC rates for phylogeography are parameters whose exact interpretation is

difficult and, as a result, estimates are rarely reported in papers. Knowing the rate at which a lineage in one location will transition to a lineage in another is of limited use when the number of lineages in the first is unknown, and the CTMC model does not estimate the latter. There is no easy way in which a prediction can be made for the risk of viral escape from that number alone. The assumed time-invariance of rates is also unlikely to apply in practice, and moreover, if a pathogen is not even present in every location at some times during the history of the tree, to nevertheless infer a rate of transition for the whole timeline raises a philosophical issue. The estimated number is then the rate at which a hypothetical lineage in an uninfected location would spread to another, but for many pathogens the very fact that the disease is present is likely to affect the rate of exit because specific measures will be put in place to control it. For example, animal exports from a Western country in the midst of an FMDV epidemic are different to exports in periods when the virus is absent. It is, as a result, not surprising that it is much more common in the literature to see links between locations identified via BSSVS, which is based on the posterior probability of a rate being nonzero and does not take into account its size if it is not. (The alternative approach, used in some analyses in both chapter 2 and chapter 4, is to reconstruct the ancestral history of the particular set of samples used in the analysis using Markov jumps [103].) So, while it is concerning that rate estimates here showed quite so much noise, it is at least reassuring that arguments are rarely made on the basis of their magnitudes. Nevertheless, this is not the only issue surrounding sampling and CTMC rates that has been uncovered; a tendency for it to overestimate rates was noted by De Maio et al. [30]. The whole CTMC discrete traits approach is prone to problems due to sampling, because it assumes that the value of each trait is something that is observed about a sequence post-sampling (like a nucleotide position) when, in fact, if it refers to is a location or host species, it is generally chosen by the investigator. An alternative is to use a structured coalescent model, which instead conditions on the subpopulation that each sample is drawn from;

the De Maio et al. paper introduces a fast approximation for the calculations needed for this purpose. Future work might apply this method to simulated data of the type presented in this chapter, to determine if the noise in estimates is reduced.

An increased sample size had much larger effect when it came to phylogeography, both for the τ statistic and the accuracy of BSSVS as a classifier. In this case, there would be a strong argument for increasing the sample size right up to 500 and, given the fitted relationships between sample size and τ, accuracy at BF=3, and ROC AUC, potentially beyond, although as all those response variables have maxima at 1 and the functions for the fitted curves do not, these relationships obviously cannot continue indefinitely. The reason for this is presumably that increasing the number of sequences increases the resolution of the transmission network; a missed sample from a particular location could easily result in an analysis which infers a non-existent link between two other locations because the true route went through the former. For example, in chapter 2, long-distance links had to be inferred, despite FMDV being transmitted largely overland, because no samples were available from intervening countries. More sequences fill in the gaps. A continuation of this study could look in detail at the effects of missing, or oversampling, a particular deme. That the BF cut-off of 3 favours specificity over sensitivity is not surprising; to be identified as well-supported, a rate that is truly zero must not merely have shown higher posterior than prior odds of being nonzero, but must obtain three times those prior odds. On the other hand, for a rate that is not zero to *fail* to meet the threshold, the posterior odds just have to fail to be three times the prior odds; it may still be more likely after the analysis than before it. It is also not surprising that a BF cut-off value of less than 1 (i.e. one that, in contrast, favours sensitivity) maximises accuracy in this simulated scenario, as most rates (all but three) were genuinely nonzero. This may genuinely be the case in the real world for some pathogens. When dealing with human pathogens spread

by air travel, for example, few rates of movement between countries are likely to be actually zero given the reach of the global aviation network. In these cases, it may actually be preferable to invert the hypothesis and test the assumption that rates are nonzero.

To conclude, I would make several recommendations as the result of this simulation study. Firstly, when selecting past sequences to use from an uneven set of samples, or designing a future study, sequence selection should be stratified by time period and location, without reference to the size of the pathogen population, or number of infections, at that location or during that period. Secondly, at least if the timescale and level of genetic diversity are comparable to the simulated situation here, sample sizes should ideally be at least 200, as even in these simple scenarios there were problems with using less. For phylogeographical analysis, there seems to be an analytical benefit in increasing sample sizes to 500 and beyond, which should be balanced against the increased computational time needed for analyses of datasets of this size, although it should not be done at the expense of unbalancing the sample with respect to location or time. Thirdly, wherever possible analysis should be repeated with different sample sets, in case any reconstructed features of the dynamics are not replicated. Fourthly, the prior distribution for the precision parameter of a skygrid (or skyride) analysis can have significant effects on the reconstruction; the default distribution in BEAST may be too diffuse and thus prefer to eliminate interesting features. At the very least, I would recommend analyses that produce reconstructions suggesting an constant population size be re-performed with a prior with a smaller median value. (In this chapter I replaced the default Gamma$(0.001, 0.001)$ prior with Gamma$(0.1, 0.1)$, which places much less prior weight on high values of the parameter.) Lastly, rate estimates from CTMC-based discrete trait analysis appear to be particularly unreliable and, as alternative methods are now available, these may be preferred.

Chapter 4

Foot-and-mouth disease virus serotype O: evolutionary history and geographical dispersal

4.1 Introduction

I now turn to another sequence analysis of foot-and-mouth-disease virus isolates, this time of serotype O, one of the two serotypes with the widest geographical distributions. It is currently present in Asia, Africa and South America, with periodic incursions to Europe [146], including the 2001 UK epidemic. Contrary to the situation with serotype SAT 2 dealt with in chapter 2, there is more serotype O data available in public databases than can be accommodated in a single phylodynamic analysis, and hence the data must be subsampled. I apply the work of chapter 3 in designing an appropriate procedure to do this.

The subdivision of the immunologically-distinct FMDV serotypes into topotypes

was proposed by Samuel and Knowles [125] in 2001 for serotype O and subsequently extended to the other serotypes [87]. The "unweighted pair group method with arithmetic mean" (UPGMA) procedure was used to divide the set of available sequences into clusters based on nucleotide similarity in the VP1 segment, with the algorithm stopped at 80% similarity for the SAT strains, and 85% for the others. The topotypes were the remaining clusters once this threshold had been reached. The initial analysis of type O, which notably used only a subsection of the VP1 gene with a length of 170 base-pairs, identified eight topotypes. Further work by Knowles et al. [84] and Ayelet et al. [5] identified three more, all from East Africa. The eleven accepted topotypes, geographically named, are as follows:

- Europe-South America (Euro-SA)
- Cathay
- South-East Asia (SEA)
- Middle East-South Asia (ME-SA)
- West Africa (WA)
- Four East African topotypes (EA-1 to EA-4)
- Two Indonesian topotypes (Indonesia-1 and Indonesia-2)

The Euro-SA topotype historically infected Europe but is now extinct there, with more recent type O epidemics in the continent being the result of incursions by others. Euro-SA was also carried to South America in the late 19th century [125] and persists there, although as of 2015 only Venezuela, Ecuador, Suriname and some areas of Brazil are not certified FMDV-free. (The list of FMDV statuses by country is maintained by the OIE and available at `http://www.oie.int/animal-health-in-the-world/official-disease-status/fmd/`). Indonesia-1

and Indonesia-2 would seem to be extinct as the country was certified FMDV-free in 1986 and no other examples of these lineages have appeared since. The Cathay topotype, which, uniquely, appears to be highly adapted to pigs (all other extant topotypes are most often recorded in cattle), frequently causes outbreaks in Hong Kong; pig farms there have been suggested as the maintenance population for the lineage [34] although the lack of data from mainland China means that the picture is unclear. It previously caused regular outbreaks in Taiwan and the Philippines, but both countries are now certified free of the virus. The remaining seven topotypes have wide ranges in countries where FMDV is endemic.

Prior to the early 1990s, several FMDV outbreaks in Europe appear to have been caused by viruses present in improperly inactivated vaccines, rather than by naturally-occurring strains [11, 146]; formaldehyde-inactivated vaccines were eventually implicated and their use was banned across the European Union in 1992, since when such events have ceased. This phenomenon has also been suspected more recently in serotype C in Kenya [128] despite the fact that formaldehyde was abandoned by the early 1980s.

As some time has elapsed since the classification of serotype O isolates into topotypes was conducted, in this chapter I repeat this exercise with a modern dataset. In addition, I perform a molecular clock analysis on VP1 segments for the entire serotype, investigating its nucleotide substitution rate and the timescale over which it has evolved; I use the random local molecular clock (RLMC) model [40] to investigate variation in mutation rates between lineages.

I then move on to investigate the phylogeography of individual topotypes. Many published sequence analyses for FMDV concentrate on isolates taken from a single country (for example [1, 28, 74, 141]); papers with an international scope are rarer, although as modern phylogeography tools have become available, they have started to appear. Di Nardo et al. [34] performed one for the Cathay topotype,

and de Carvalho et al. [29], in concentrating on South America, implicitly did the same for Euro-SA. Of the remaining topotypes, I fill in some of the remaining gaps by investigating two more, SEA and ME-SA. In addition to reconstructing spatial movements, I use the recently-developed general linear model (GLM) method [45, 93] to identify characteristics of countries that are predictors of FMDV lineage movement between them.

4.2 Methods

All records (as of January 2015) for serotype O isolates of FMDV that included at least 95% of the VP1 segment were downloaded from the NCBI Nucleotide database. These were then deduplicated, and records for strains that were not field isolates, had no given dates, or were isolates from outbreaks whose origins were confirmed as, or thought likely to be, laboratory escapes [146] were excluded. The remaining sequences were aligned using MAFFT [79], and then trimmed to the VP1 segment only. The process used by Samuel and Knowles [125] in identifying type O topotypes by clustering isolates using UPGMA and a threshold of 85% nucleotide similarity on VP1 was repeated using this full alignment.

An unrooted neighbour-joining tree of the complete dataset, and one for each identified cluster, were constructed using the TN93 distance matrix [142]. These, along with the sampling date of each isolate, were used as input for the program Path-O-Gen (`http://tree.bio.ed.ac.uk/software/pathogen/`) in order to investigate the strength of the molecular clock signal in the dataset. Path-O-Gen calculates, for a given rooted tree with branch lengths in units of nucleotide substitutions, the Pearson product-moment correlation coefficient of the date of sampling and the root-to-tip divergence (RTTD) of each isolate; the latter is the distance along the tree branches from the root to the tip representing that isolate.

It also identifies the root position that maximises this correlation coefficient. With this root identified, a model of RTTD as a function of sampling date was fit using simple linear regression.

All subsequent analyses relied upon taking a smaller sample from the full alignment according to one stochastic sampling scheme or another, and in every case, this was conducted ten times and the sequence data making up each replicate of the scheme analysed separately. Firstly, a molecular clock analysis was conducted for the entire serotype by downsampling it to include only one isolate per country per five-year time period (with breaks at the beginning and middle of each decade). The resulting alignment was analysed using BEAST 1.8.1 [39], with the RLMC as a clock model [40], the SRD06 nucleotide substitution model [132] and a Bayesian skygrid tree prior [52]. The midpoint of the year of sampling of each isolate from the NCBI database was used as a tip date. The RLMC was chosen as the possibility that substitution rates vary by topotype has been suggested in the literature [34]. In contrast to the more commonly employed uncorrelated relaxed clock models, which assume that branch-specific substitution rates are drawn from a single probability distribution and there is no correlation between the rates on adjacent branches, the RLMC assumes that rates change only at a small number of positions in the tree, with most branches having exactly the same rate as their neighbours. Mixing of the Markov chain for the RLMC is unfortunately much poorer than for uncorrelated models, and as a result I employed Metropolis-coupled MCMC in BEAST, with a total of four chains, two unheated and two only slightly heated with temperatures of 0.99 and 0.98. Chains were run for long enough to ensure ESSs of at least 100 for all numerical parameters.

The SEA and ME-SA topotypes were then used for a phylogeography analysis. Only sequences for isolates collected from 1995 onwards were considered, and sequences were also excluded if less than five examples from their country of origin were available over this time period, or the country of origin was certified wholly

FMDV-free at any point during it (this was taken as representing a country in which FMDV is nowhere endemic). A number of sequences which the full-serotype molecular clock analysis had suggested were most likely from clinical cases caused by inadequately inactivated vaccines, or the descendants of them, rather than naturally occurring viruses, were also excluded. A random selection was then obtained using an algorithm intended to ensure that similar numbers of sequences were analysed from each country, and that, for each country, the selected sequences were spread out as widely as possible over time. For each country, this proceeded as follows, starting with a pool P of sequences sampled in that country:

1. Empty the set S.
2. Empty the set Y.
3. Randomly select an odd-numbered year y between 1995 and 2013, that is not in Y. If there is no such year, go to 6).
4. Add y to Y.
5. If there are any remaining sequences in P isolated in the two-year period starting with y, select one, remove it from P, and add it to S.
6. If P is empty, stop.
7. If S has ten or more elements, stop.
8. Go to 2).

The algorithm picks 10 sequences, or all the sequences that are available from a given country if this is less than 10 (remembering that only countries for which at least five sequences were available was included). It also ensures, as far as possible, that each two-year period from 1995 to 2014 is represented in the sample.

The phylogeography analysis was conducted using a GLM predictor model [45, 93]. The structure of this model is as follows: suppose $\mathbf{\Lambda}$ is the matrix determining the rate at which lineages move between countries, such that an entry Λ_{ij}, $i \neq j$, is the rate of transition between country i and country j, and $\mathbf{\Lambda}$ is normalised such that one transition occurs per unit time. A predictor is a (possibly asymmetric) relationship between countries, taking a numerical value, that may influence the overall rate of movement. Suppose we have N of them; each defines a predictor matrix $\mathbf{x}_k$ whose entries $x_{i,j,k}$ are the values of the predictor when i is the source country and j the destination. The model then assumes that, for all i, j, $i \neq j$.

$$\ln \Lambda_{ij} = \sum_{k=1}^{n} \beta_k \delta_k x_{i,j,k}$$

The β_ks are coefficients that can take any real value; as such movement rates may be either positively or negatively correlated with predictor values. The δ_ks are indicator variables; $\delta_k = 1$ if $\mathbf{x}_k$ is included in the model, or 0 if it is not. A prior distribution is placed on each β_k and on the total number of predictors that are included.

For geographical diffusion of FMDV, I used a total of 26 predictors. 22 of these were derived from data about the countries and their relationships with each other, while the last four concerned the quantity and temporal distribution of the available samples and were intended to control for any effect that these might have on the analysis:

- The population of cattle, goats, pigs, sheep, and water buffalo (*Bubalus bubalis*) in the source country.

- The population of cattle, goats, pigs, sheep, and *B. bubalis* in the destination country.

- The amount of trade in live cattle, goats, pigs, sheep, and *B. bubalis* from the source country to the destination country.
- The amount of trade, in tonnes, of whole fresh cow milk, cattle meat, goat meat, pig meat and sheep meat from the source country to the destination country.
- The minimum spatial distance in kilometers between the two countries.
- The minimum number of land borders separating the two countries (intended to be a measure of the "bureaucratic distance" involved in land travel).
- The number of sequences included in the analysis from the source country.
- The number of sequences included in the analysis from the destination country.
- The number of two-year time periods, 1995-2014, which provided a sequence in the sample from the source country.
- The number of two-year time periods, 1995-2014, which provided a sequence in the sample from the destination country.

The data for the first twenty predictors were obtained from the FAOSTAT database. For animal populations, figures used were the mean population size recorded from 1995 to 2011 (which was the last year for which data was available). Trade matrices in FAOSTAT are, unfortunately, not comprehensive; only some countries report their imports and exports by partner country. It is therefore possible to quantify trade from a reporting country to any other country, and vice versa, but not possible between two non-reporting countries. Those values in the predictor matrices had to be taken to be zero. In addition, figures for trade between reporting countries are not always consistent; FAOSTAT may have the source country reporting a number of exported animals to the destination country and

the destination reporting a *different* number of imported animals from the source. As a result, for each year from 1995 to 2011, the amount of trade between each pair of countries (in both directions) was taken to be, if both reported a nonzero figure, the mean of those two figures, or if only one did, that figure. The mean of these numbers over the 17-year period was then used as the predictor. A pseudocount of 1 was added to any predictor whose actual value was 0, and then all values were log-transformed and scaled so the total set of values for each predictor had a mean of 0 and a standard deviation of 1. For the sets of countries of origin for both topotypes, correlation between predictors was investigated by calculating the Pearson product-moment correlation coefficient between each pair of normalised predictor values. The prior distribution for each β_k was normal with a mean of 0 and a standard deviation of 2. The prior assumption regarding the set of δ_ks was that each was a Bernoulli trial with success probability 0.0263; this gives a 50% prior probability that no predictors are included at all. As the prior odds for inclusion of a single predictor are therefore 0.027, the posterior odds can be used to calculate a Bayes Factor for the additional support that the data provide for each predictor's inclusion in the model.

The BEAST analysis used a skygrid tree prior, SRD06 substitution model, and uncorrelated lognormal relaxed molecular clock. Convergence and mixing of the MCMC chain is much better for uncorrelated clock models than for the RLMC, so I did not employ Metropolis-coupled MCMC and all chains were run for long enough to ensure ESSs of at least 200 for all numerical parameters. BSSVS is utilised here to identify relevant predictors of movement, rather than to identify nonzero rates of movement between particular locations as it was in chapter 2. The consequence of this is that Bayes Factor tests for individual between-country rates cannot be employed. Instead, the Markov Jumps procedure [103] was used to reconstruct, for every sampled tree from the MCMC, the history of lineage transitions between countries. This was summarised over the full posterior

distribution by, for each pair of countries in the analysis, calculating the proportion of samples that reconstructed at least one jump from one to the other (in both directions), and the median number of such jumps. Between-country links for which there was at least a 90% posterior probability that one such jump occurred were considered to be well-supported by the data.

Another discrete traits analysis was conducted for ME-SA and SEA using host species as the trait. Due to lack of data from other species, only sequences from cattle, pigs, *B. bubalis* and, for ME-SA only, sheep were included. Sets of sequences were selected using the same sampling procedure employed for countries of origin, except that a maximum of thirty samples per species were included. The historical numbers of host jumps were again reconstructed using Markov jumps.

4.3 Results

4.3.1 Analysis of the full serotype

A total of 1816 VP1 segments from unique isolates were retrieved from the database. When the UPGMA procedure originally used to identify topotypes [125] was repeated on the current NCBI data, twelve clusters were identified. Figure 4.1 depicts an unrooted neighbour-joining topology with branches coloured by these twelve groups. Notably, of the eleven accepted topotypes in the current literature [5, 84, 125], EA-1 and EA-3 are not distinct and the combined cluster including them is not monophyletic, ME-SA is also not monophyletic, and the original Euro-SA topotype is split into three, which I designate as (1), (2) and (3). While Euro-SA (2) consists only of sequences from isolates sampled prior to 1967, the other two both contain examples dating from 2010 or later. I refer to these groups

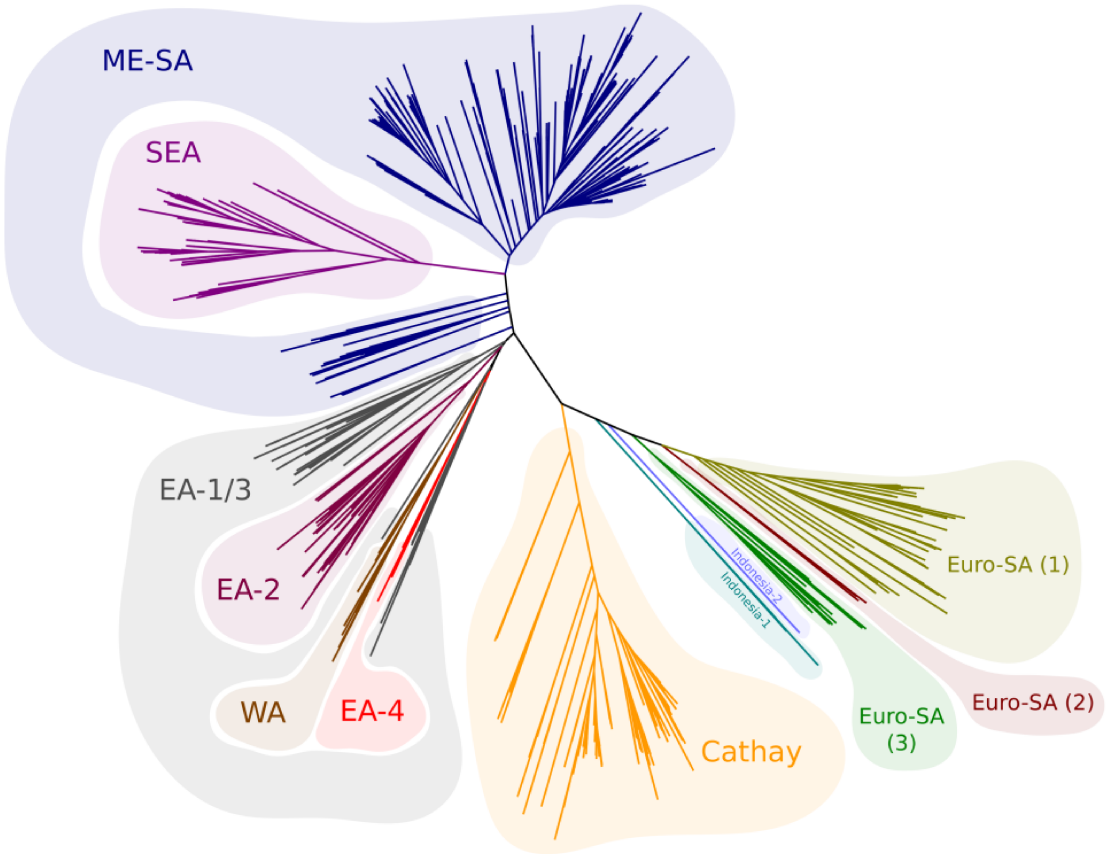


Figure 4.1: Unrooted neighbour-joining phylogeny for every FMDV serotype O VP1 sequence in the NCBI Nucleotide database, calculated using the TN93 distance matrix. Branches are coloured and annotated according to topotype clusters assigned by the UPGMA algorithm with a threshold of 0.85 nucleotide similarity.

as "topotype clusters" henceforth, to clarify that this classification refers to the results of my analysis, not the standard set of topotypes from the literature.

Figure 4.2 plots RTTD against sampling date for the tree in figure 4.1, with the root positioned such that the Pearson product-moment correlation coefficient is maximised. This maximum value is 0.346; while a positive relationship is obvious ($p < 2.2 \times 10^{-16}$ against the hypothesis that date and RTTD are not associated), the correlation is weak. The points in the figure are coloured by the

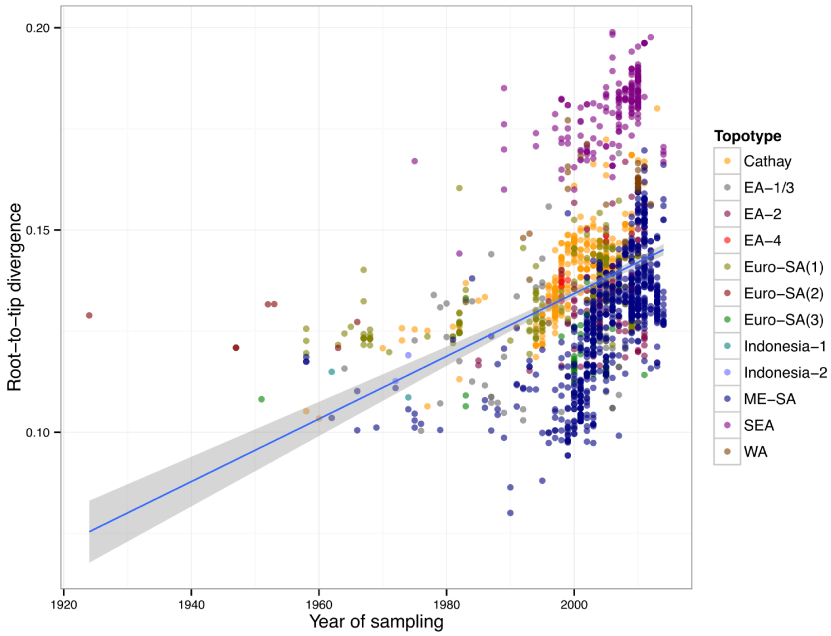


Figure 4.2: Scatter plot of root-to-tip divergence (using the neighbour-joining phylogeny from figure 4.1, with the best-fit root identified by Path-O-Gen) versus year of sampling for all 1816 serotype O VP1 sequences. The blue line was fit by simple linear regression; the grey area represents the 95% confidence interval for the slope of this line. Points are coloured by topotype cluster.

topotype clusters identified by UPGMA, and examination of those by eye suggests within-cluster relationships that are not captured by the overall regression line.

In fact, if the sequences are separated into topotype clusters and the same procedure conducted for each (figure 4.3), high correlation coefficients are obtained for all except Euro-SA(3) and WA, and clear evidence for an association between date and RTTD for all except WA. (The very small number of available sequences for the two Indonesian clusters meant they had to be excluded here.) This suggests that mutation in serotype O does indeed exhibit clock-like behaviour, but that variation between lineages, or over time, results in a complete picture that deviates strongly from that which would be expected under a strict clock.

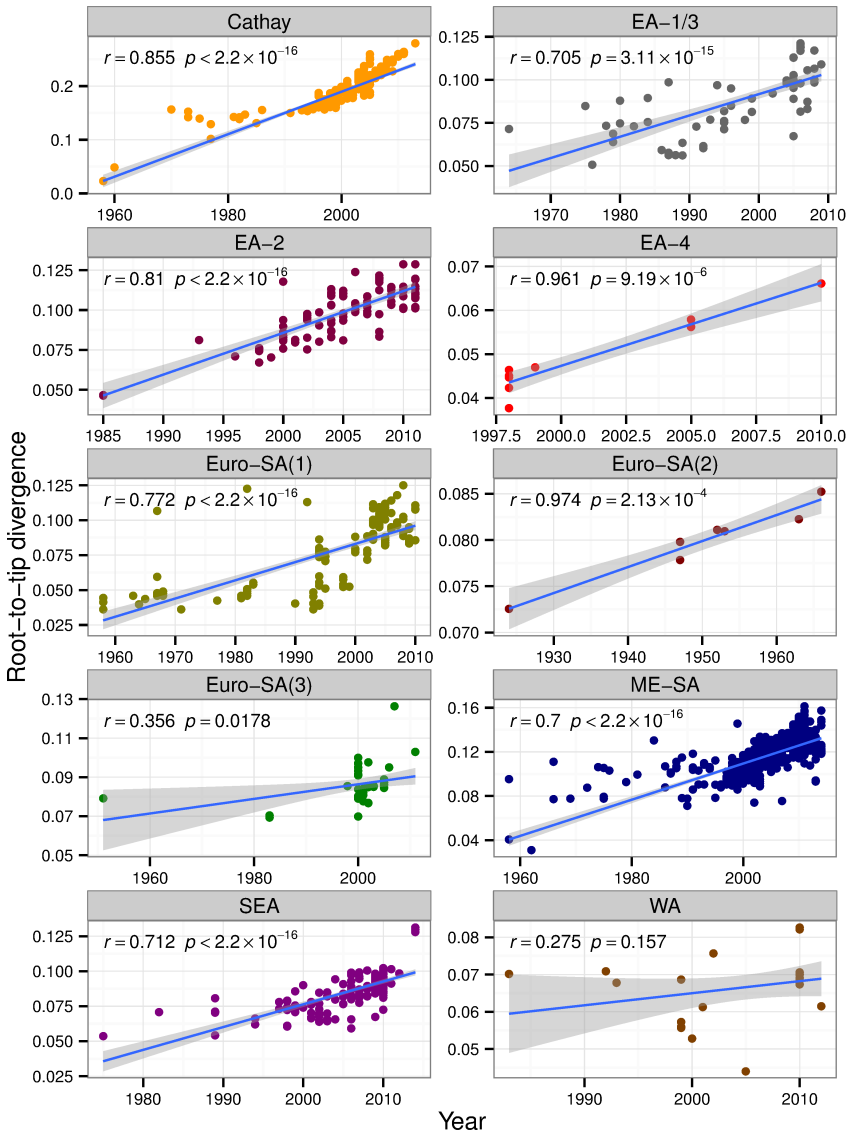


Figure 4.3: Scatter plot of root-to-tip divergence (using neighbour-joining phylogenies generated with the TN93 distance matrix, with the best-fit root identified by Path-O-Gen) versus year of sampling for the sequences comprising 10 of the 12 topotype clusters. Blue lines were fit by simple linear regression; the grey area represents the 95% confidence interval for the slope of this line. Each plot is annotated with the Pearson product-moment correlation coefficient and the p-value for the hypothesis test.

While I analysed ten replicates of each sampling scheme in every case, here in general I present only one of each; notable between-replicate variation is mentioned in the text. More comprehensive results of all ten are given in appendix B. A sampling scheme of picking one available sequence per country per five-year period resulted in a dataset of 233 sequences. Figure 4.4 is the maximum clade credibility (MCC) tree, with the branches coloured by topotype cluster. The non-monophyletic nature of ME-SA and EA-1/3 is confirmed here, and the three clusters making up the traditional Euro-SA topotype have a very early TMRCA, in March 1905 (95% HPD: March 1896-February 1913); this is also the TMRCA of the two extant clusters. The TMRCA of the entire serotype has a posterior median of July 1894 (February 1877-May 1905). All but one sampling replicate gave results similar to this for the TMRCA; the lone exception having a much earlier posterior median of May 1866 (March 1851-May 1881). The posterior support for the clade descended from the MRCA of each cluster was 1 for all except ME-SA and Euro-SA(3), which both had a probability of 0.987.

Figure 4.5 is the same tree, except that branches are now coloured by posterior median substitution rate. Tips are annotated by topotype cluster and by the posterior median and 95% HPD for the substitution rate on the terminal branch leading to that tip; these rates are in units of 0.001 substitutions per site per year. The posterior distribution for RLMC output is very difficult to summarise succinctly, but notable features are:

- Over most of the tree, the posterior median rate is around 4×10^{-3} substitutions per site per year.
- The clade consisting of almost all ME-SA isolates samplied since 2000 has a faster median rate of around 5.5×10^{-3}, but some of the older isolates in that topotype show much slower numbers than this, of around 2.6×10^{-3}. It should be noted that only eight out of ten sampling replicates showed this

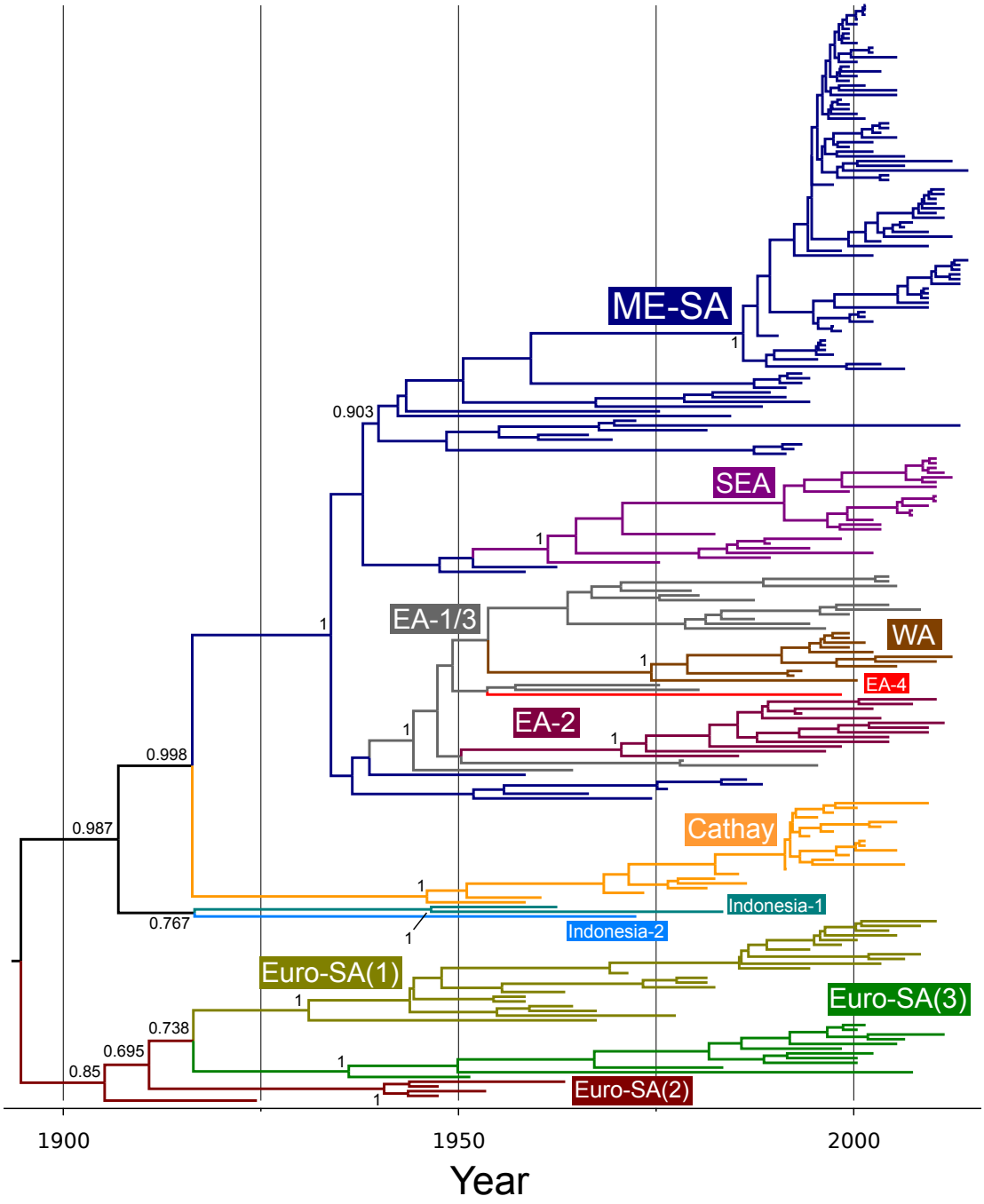
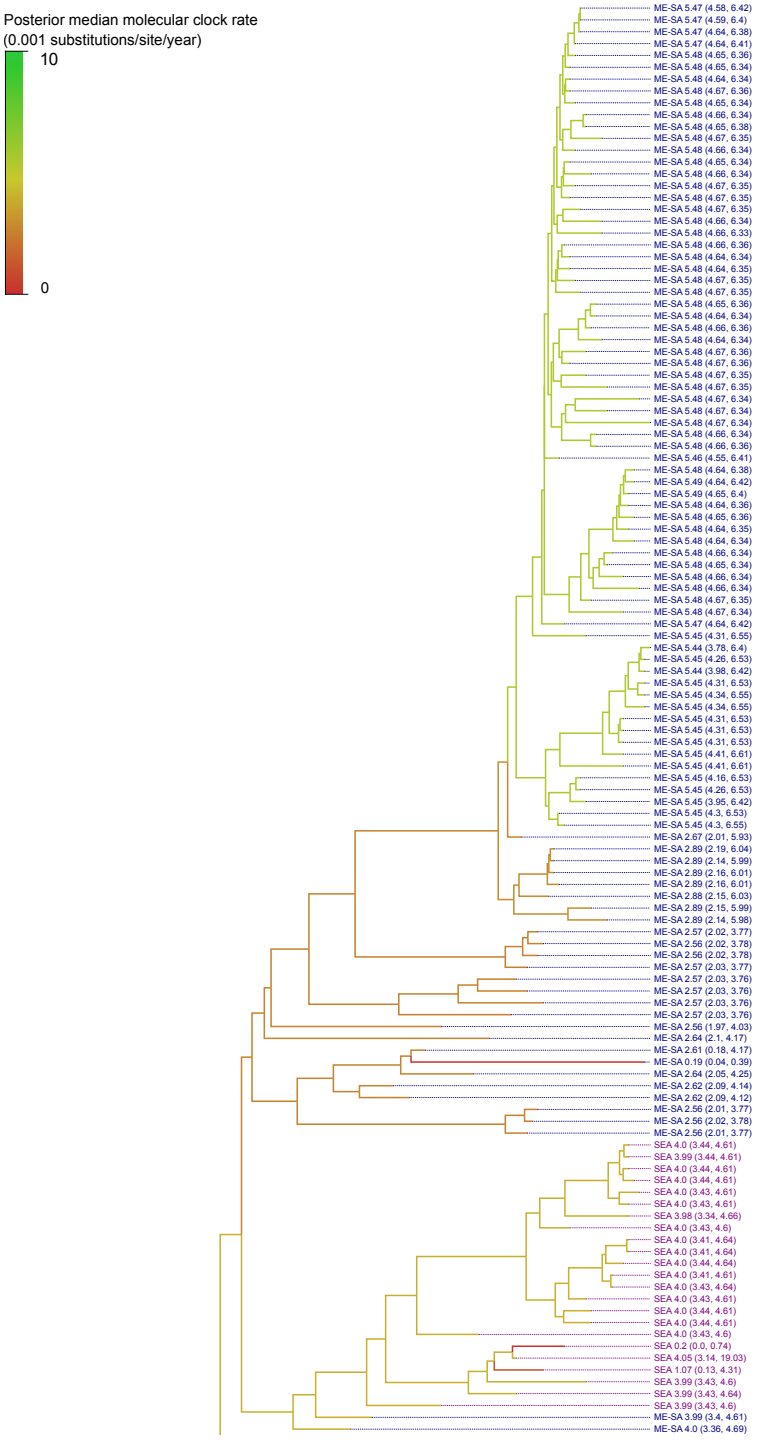


Figure 4.4: Maximum clade credibility phylogeny of an analysis of 233 serotype O sequences, sampled such that one was selected from every available country and five-year time period. Branches are coloured by topotype cluster as in figure 4.1. Selected nodes of each topotype are annotated with posterior probabilities for the existence of the descendant clade.

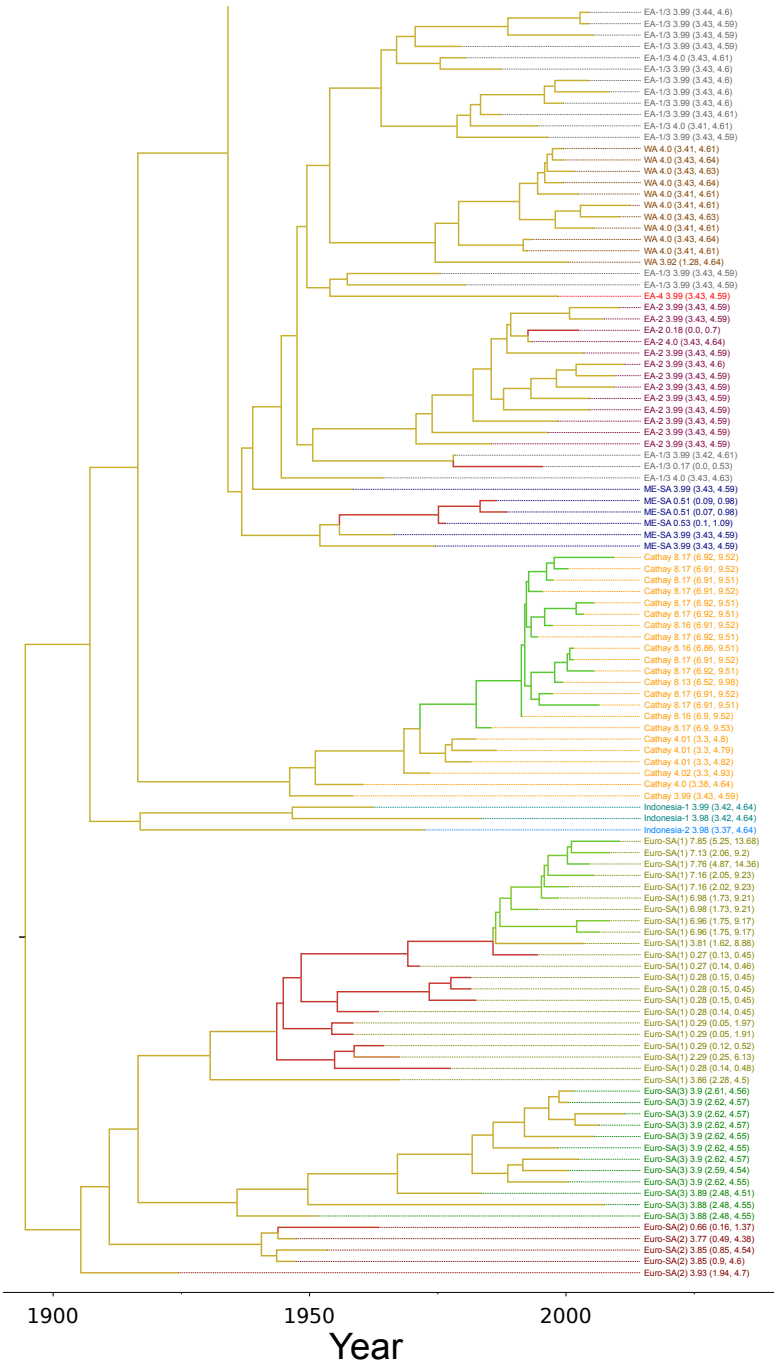
split. The remaining two assigned similar rates to the great majority of the ME-SA branches as to rest of the tree (i.e. a median of around 4×10^{-3}).

- Recent (since around 1990) Cathay branches have a very fast median rate of around 8.15×10^{-3}.
- Mid-century Euro-SA(1) branches have an unfeasibly slow median rate of around 3×10^{-4}. Most later isolates are assigned fast median rates of around 7×10^{-3}, but most of those estimates have wide HPD intervals whose lower bounds are less than 2×10^{-3}.
- A number of individual tips exist with unusually, and again unfeasibly, slow rates of less than 3×10^{-4}.

The tree illustrated in figure 4.5 illustrates one common configuration of tree topology and branch rates for Euro-SA(1), with the later tips all being descendants of lineages with very slow rates. All the tips with those slow rates, which correspond to isolates from six countries on both continents sampled over a period of 36 years, share at least 98% nucleotide similarity on VP1 with the widely used vaccine strain O_1 Campos/58 [22, 89, 90, 99], and are almost certainly the result of outbreaks caused by inadequately inactivated vaccines; the apparent lack of mutation is in fact due to repeated human reintroductions of that 1958 strain. In the scenario presented in figure 4.5, many contemporary South American Euro-SA lineages are in fact the descendants of the viruses in those vaccines. However, there is another configuration which is also frequent in the posterior distribution, in which the slow branches form a clade on their own, and a path from the contemporary strains to the root does not encounter a branch with such a rate; in this scenario the latter have no vaccine ancestry. Figure 4.7 illustrates this. The posterior probability that *all* Euro-SA(1) isolates from 1990 onwards that are not themselves very similar to O_1 Campos/58 have ancestral branches with clock rates less than 10^{-4} substitutions per site per year is 0.362, the probability that some, but not all do


Posterior median molecular clock rate
(0.001 substitutions/site/year)
10
0
ME-SA 5.47 (4.58, 6.42)
ME-SA 5.47 (4.59, 6.4)
ME-SA 5.47 (4.64, 6.38)
ME-SA 5.47 (4.64, 6.41)
ME-SA 5.48 (4.65, 6.36)
ME-SA 5.48 (4.65, 6.34)
ME-SA 5.48 (4.64, 6.34)
ME-SA 5.48 (4.67, 6.36)
ME-SA 5.48 (4.65, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.65, 6.38)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.65, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.66, 6.33)
ME-SA 5.48 (4.66, 6.36)
ME-SA 5.48 (4.64, 6.34)
ME-SA 5.48 (4.64, 6.35)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.65, 6.36)
ME-SA 5.48 (4.64, 6.34)
ME-SA 5.48 (4.66, 6.36)
ME-SA 5.48 (4.64, 6.34)
ME-SA 5.48 (4.67, 6.36)
ME-SA 5.48 (4.67, 6.36)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.34)
ME-SA 5.48 (4.67, 6.34)
ME-SA 5.48 (4.67, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.66, 6.36)
ME-SA 5.46 (4.55, 6.41)
ME-SA 5.48 (4.64, 6.38)
ME-SA 5.49 (4.64, 6.42)
ME-SA 5.49 (4.65, 6.4)
ME-SA 5.48 (4.64, 6.36)
ME-SA 5.48 (4.65, 6.36)
ME-SA 5.48 (4.64, 6.35)
ME-SA 5.48 (4.64, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.65, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.66, 6.34)
ME-SA 5.48 (4.67, 6.35)
ME-SA 5.48 (4.67, 6.34)
ME-SA 5.47 (4.64, 6.42)
ME-SA 5.45 (4.31, 6.55)
ME-SA 5.44 (3.78, 6.4)
ME-SA 5.45 (4.26, 6.53)
ME-SA 5.44 (3.98, 6.42)
ME-SA 5.45 (4.31, 6.53)
ME-SA 5.45 (4.34, 6.55)
ME-SA 5.45 (4.34, 6.55)
ME-SA 5.45 (4.31, 6.53)
ME-SA 5.45 (4.31, 6.53)
ME-SA 5.45 (4.31, 6.53)
ME-SA 5.45 (4.41, 6.61)
ME-SA 5.45 (4.41, 6.61)
ME-SA 5.45 (4.16, 6.53)
ME-SA 5.45 (4.26, 6.53)
ME-SA 5.45 (3.95, 6.42)
ME-SA 5.45 (4.3, 6.53)
ME-SA 5.45 (4.3, 6.55)
ME-SA 2.67 (2.01, 5.93)
ME-SA 2.89 (2.19, 6.04)
ME-SA 2.89 (2.14, 5.99)
ME-SA 2.89 (2.16, 6.01)
ME-SA 2.89 (2.16, 6.01)
ME-SA 2.88 (2.15, 6.03)
ME-SA 2.89 (2.15, 5.99)
ME-SA 2.89 (2.14, 5.98)
ME-SA 2.57 (2.02, 3.77)
ME-SA 2.56 (2.02, 3.78)
ME-SA 2.56 (2.02, 3.78)
ME-SA 2.57 (2.03, 3.77)
ME-SA 2.57 (2.03, 3.76)
ME-SA 2.57 (2.03, 3.76)
ME-SA 2.57 (2.03, 3.76)
ME-SA 2.57 (2.03, 3.76)
ME-SA 2.56 (1.97, 4.03)
ME-SA 2.64 (2.1, 4.17)
ME-SA 2.61 (0.18, 4.17)
ME-SA 0.19 (0.04, 0.39)
ME-SA 2.64 (2.05, 4.25)
ME-SA 2.62 (2.09, 4.14)
ME-SA 2.62 (2.09, 4.12)
ME-SA 2.56 (2.01, 3.77)
ME-SA 2.56 (2.02, 3.78)
ME-SA 2.56 (2.01, 3.77)
SEA 4.0 (3.44, 4.61)
SEA 3.99 (3.44, 4.61)
SEA 4.0 (3.44, 4.61)
SEA 4.0 (3.44, 4.61)
SEA 4.0 (3.43, 4.61)
SEA 4.0 (3.43, 4.61)
SEA 3.98 (3.34, 4.66)
SEA 4.0 (3.43, 4.6)
SEA 4.0 (3.41, 4.64)
SEA 4.0 (3.41, 4.64)
SEA 4.0 (3.44, 4.64)
SEA 4.0 (3.41, 4.61)
SEA 4.0 (3.43, 4.64)
SEA 4.0 (3.43, 4.61)
SEA 4.0 (3.44, 4.61)
SEA 4.0 (3.44, 4.61)
SEA 4.0 (3.43, 4.6)
SEA 0.2 (0.0, 0.74)
SEA 4.05 (3.14, 19.03)
SEA 1.07 (0.13, 4.31)
SEA 3.99 (3.43, 4.6)
SEA 3.99 (3.43, 4.64)
SEA 3.99 (3.43, 4.6)
ME-SA 3.99 (3.4, 4.61)
ME-SA 4.0 (3.36, 4.69)

(a) (Top)



(b) (Bottom)

Figure 4.5: Maximum clade credibility phylogeny of an analysis of 233 serotype O sequences. Branches are coloured by posterior median molecular clock rate; tips are labelled with topotype clusters and the posterior median and 95% HPD interval for that rate on the terminal branch leading to that tip.

is 0.412, and the probability than none do is 0.226. (The scenario in which some, but not all, isolates have such an ancestry was the most probable in nine out of ten sampling replicates; in one the scenario that none did was preferred with a posterior probability of 0.715.) To establish that this finding was not the result of the use of a molecular clock model on a part of the phylogeny that has clear reasons for not exhibiting clock-like behaviour, I also performed an analysis of the full collection of Euro-SA(1) sequences using a non-clock model in MrBayes 3.2 [121]. This gave a posterior probability of 0.964 for at least one later sequence lying in the clade whose MRCA is that of all the Campos-like sequences.

The reconstructed skygrid plot is shown in figure 4.6. The median line suggests that the effective population size peaked around 1950 and then subsequently experienced a slow decline, arrested by another peak after 2000. However, the width of the HPD intervals is such that a population size undergoing little variation between around 1950 and 2005 is also consistent with the data.

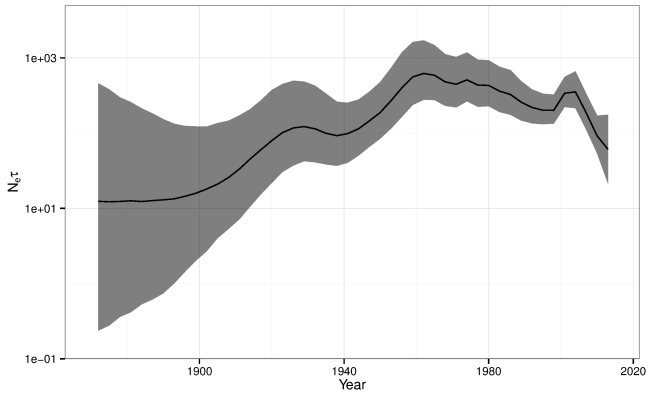


Figure 4.6: Reconstructed skygrid plot from analysis of the full serotype. The black line follows variation in the the posterior median population size through time, and the grey area marks the 95% HPD intervals.

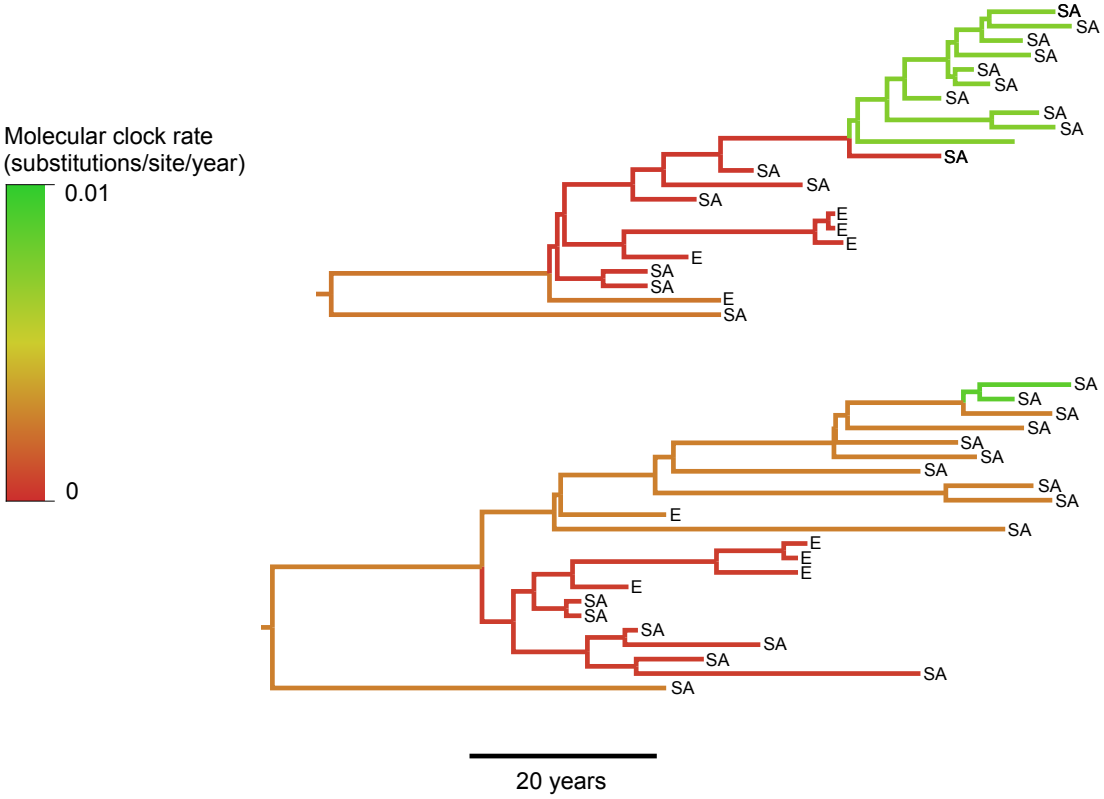


Figure 4.7: Two possible ancestral scenarios for the Euro-SA(1) clade in the RLMC analysis. Both are enlargements of regions of actual trees from the posterior distribution. Branches are coloured by molecular clock rate. Tips are labelled by continent of sampling (E=Europe, SA=South America). Top: recent South American isolates with fast clock rates on the terminal branches are descended from branches with very slow clock rates. Bottom: Scenario 2, isolates with very slow clock rates form a distinct clade.

4.3.2 Topotype SEA

The SEA cluster comprised 200 sequences. Of these, seven came from isolates sampled before 1995, six had a country of origin that provided less than five sequences in total (Cambodia, the Democratic People's Republic of Korea, and Russia) and ten were from single outbreaks in Japan, the Republic of Korea and Taiwan, where FMDV is not endemic. The size of the total pool of sequences for the GLM phylogeography analysis was therefore 177, of which each replicate of the country-based sampling scheme selected 78, from a total of eight countries (treating mainland China and the Hong Kong Special Administrative Region separately). See table 4.2 for a list of these. Three countries did not provide detailed import or export data to FAOSTAT and, as a result, trade between those three was recorded as zero in the predictor matrices. Table 4.1 gives the Pearson correlation coefficient between every pair of GLM predictors. Predictors for which there was a strong correlation (absolute value of coefficient ≥ 0.7) were as follows:

- cattle population and goat population (at either origin or destination)
- cattle population and pig population (at either origin or destination)
- trade in live cattle and trade in live goats
- trade in cow milk and trade in cattle meat
- trade in cow milk and trade in pig meat
- trade in cattle meat and trade in pig meat
- trade in cattle meat and trade in sheep meat
- minimum spatial distance and number of intervening land borders
- number of sequences and number of two-year sampling periods represented (at either origin or destination)

Predictor	OBP	OCP	OGP	OPP	OSP	DBP	DCP	DGP	DPP	DSP	TB	TC	TG	TP	TS	TCMi	TCMe	TGMe	TPMe	TSMe	MSD	ILB	OSC	DSC	OSP
OCP	0.54																								
OGP	0.26	**0.92**																							
OPP	**0.91**	0.64	0.46																						
OSP	-0.1	0.53	0.67	0.12																					
DBP	-0.14	-0.08	-0.04	-0.13	0.01																				
DCP	-0.08	-0.14	-0.13	-0.09	-0.08	0.54																			
DGP	-0.04	-0.13	-0.14	-0.07	-0.1	0.26	**0.92**																		
DPP	-0.13	-0.09	-0.07	-0.14	-0.02	**0.91**	0.64	0.46																	
DSP	0.01	-0.08	-0.1	-0.02	-0.14	-0.1	0.53	0.67	0.12																
TB	0.18	0.13	0.05	0.12	0	0.13	0.05	-0.08	0.05	0.04															
TC	0.38	0.34	0.21	0.4	0.23	0.07	-0.08	-0.16	-0.02	-0.02	0.64														
TG	0.28	0.32	0.28	0.35	0.32	0	-0.18	-0.24	-0.07	-0.03	0.46	**0.71**													
TP	0.31	0.28	0.2	0.41	0.25	-0.01	-0.2	-0.18	0	-0.15	0.01	0.54	0.43												
TS	0.26	0.26	0.25	0.34	0.3	0.01	-0.2	-0.27	-0.07	-0.06	0.55	0.62	0.59	0.3											
TCMi	0.14	-0.03	-0.06	0.22	0.16	-0.11	-0.28	-0.22	-0.04	-0.1	-0.01	0.37	0.3	0.65	0.31										
TCMe	0.04	-0.1	-0.11	0.11	0.04	0.05	-0.14	-0.11	0.09	-0.05	0.1	0.36	0.24	0.36	0.39	**0.74**									
TGMe	0.04	0.14	0.24	0.21	0.28	0.03	-0.09	-0.07	0.09	0.07	-0.03	0.31	0.44	0.35	0.53	0.43	0.53								
TPMe	0.12	-0.09	-0.12	0.21	-0.08	-0.08	-0.32	-0.28	-0.03	-0.15	0.05	0.29	0.22	0.44	0.29	**0.75**	**0.8**	0.47							
TSMe	-0.05	-0.14	-0.06	0.08	0.14	0.06	-0.15	-0.14	0.13	-0.07	-0.01	0.21	0.3	0.33	0.44	0.69	**0.75**	**0.73**	0.69						
MSD	-0.41	-0.31	-0.22	-0.41	-0.03	-0.41	-0.31	-0.22	-0.41	-0.03	-0.31	-0.45	-0.31	-0.43	-0.15	-0.22	-0.15	-0.21	-0.12	-0.17					
ILB	-0.31	-0.28	-0.2	-0.31	0	-0.31	-0.28	-0.2	-0.31	0	-0.27	-0.37	-0.28	-0.33	-0.05	-0.08	-0.02	-0.16	-0.03	-0.08	**0.91**				
OSC	-0.38	-0.5	-0.6	-0.67	-0.53	0.05	0.07	0.09	0.1	0.08	-0.01	-0.31	-0.37	-0.38	-0.4	-0.25	-0.13	-0.46	-0.19	-0.3	0.3	0.25			
DSC	0.05	0.07	0.09	0.1	0.08	-0.38	-0.5	-0.6	-0.67	-0.53	0.12	0.15	0.14	-0.06	0.15	-0.06	-0.12	-0.16	-0.02	-0.12	0.3	0.25	-0.14		
OSP	0.3	0	-0.28	-0.09	-0.42	-0.04	0	0.04	0.01	0.06	0.14	-0.05	-0.14	-0.2	-0.19	-0.21	-0.18	-0.43	-0.23	-0.39	0.01	0.01	**0.72**	-0.1	
DSP	-0.04	0	0.04	0.01	0.06	0.3	0	-0.28	-0.09	-0.42	0.2	0.21	0.12	-0.11	0.13	-0.22	-0.13	-0.18	-0.17	-0.17	0.01	0.01	-0.1	**0.72**	-0.14

Table 4.1: Pairwise Pearson product-moment correlation coefficients for predictors of movement between countries included in the SEA analysis. Coefficients with absolute values greater than 0.7 are in bold type. OBP, OCP, OGP, OPP, OSP: buffalo, cattle, goat, pig and sheep populations at origin. DBP, DCP, DGP, DPP, DSP: populations at destination. TB, TC, TG, TP, TS: trade in live animals. TCMi: trade in cow milk. TCMe, TGMe, TPMe, TSMe: trade in cattle, goat, pig and sheep meat. MSD: minimum spatial distance. ILB: intervening land borders. OSC, DSC: sample count from origin and destination. OSP, DSP: 2-year sampling periods represented at origin and destination. Figures are rounded to two decimal places.

In the GLM analysis, cattle trade is supported as predictor of movement with BF=19575.16; the coefficient implies the intuitive positive relationship between volume of trade and FMDV lineage movement. No other predictor obtained BF support of more than 3. Indeed, none had support of more than 1: the analysis favours their exclusion. Figure 4.8 depicts the full results.

The posterior median TMRCA for the SEA sequences sampled after 1995 was August 1986 (October 1979-January 1992). This is not the TMRCA of the entire cluster as identified above; in figure 4.1, some earlier sequences are basal to the clade containing tips from this time period. This date is notably much later than the estimated TMRCA of the latter clade in the analysis of the full serotype, which was November 1964 (April 1960-February 1966), and the HPDs do not overlap. The reason for this can be seen in the fact that the molecular clock for the entire tree in this analysis was estimated to be faster than the rates assigned to the equivalent branches in the earlier analysis, with a mean of 8.1×10^{-3} substitutions per site per year ($6.01 \times 10^{-3} - 0.0107$) and a standard deviation of 5.34×10^{-3} ($2.67 \times 10^{-3} - 9.18 \times 10^{-3}$). (Normal BEAST output, counter-intuitively, gives the mean of the lognormal distribution of clock rates on the real scale and the standard deviation on the log scale, and these are the numbers that are generally reported in papers. I depart from this and give both on the real scale.) A lognormal distribution with these parameters has a median of 6.75×10^{-3}.

Figure 4.9 is the MCC phylogeny; tips are labelled by, and branches are coloured by, country. The tree is annotated according to specific SEA lineages (Mya-98 and Cam-94) that have previously been identified in the literature [1, 82]. Myanmar had the highest posterior probability (0.619) for the location of the root node, followed by Thailand (0.255). No other country had a probability above 0.1.

Figure 4.10 displays the reconstructed skygrid plot for SEA only. A decline in effective population size from around 2002 can be observed. Figure 4.11 summarises

SEA		ME-SA	
Country	Count	Country	Count
China (mainland)	8	**Afghanistan**	12
China (Hong Kong SAR)	18	**Bangladesh**	21
Laos	16	**Bhutan**	35
Malaysia	25	**Cambodia**	6
Mongolia	14	**China (mainland)**	9
Myanmar	24	**Egypt**	5
Thailand	43	**India**	188
Viet Nam	29	**Iran**	26
		Israel	23
		Kazakhstan	6
		Laos	21
		Libya	13
		Malaysia	19
		Nepal	42
		Pakistan	100
		Saudi Arabia	14
		Turkey	56
		United Arab Emirates	11
		Viet Nam	30
Total	177	Total	737

Table 4.2: Countries from which sequences were included in the SEA and ME-SA topotype analyses, and the total number of sequences included in the full poll for each. Those in bold type reported by-country import and export statistics to FAOSTAT.

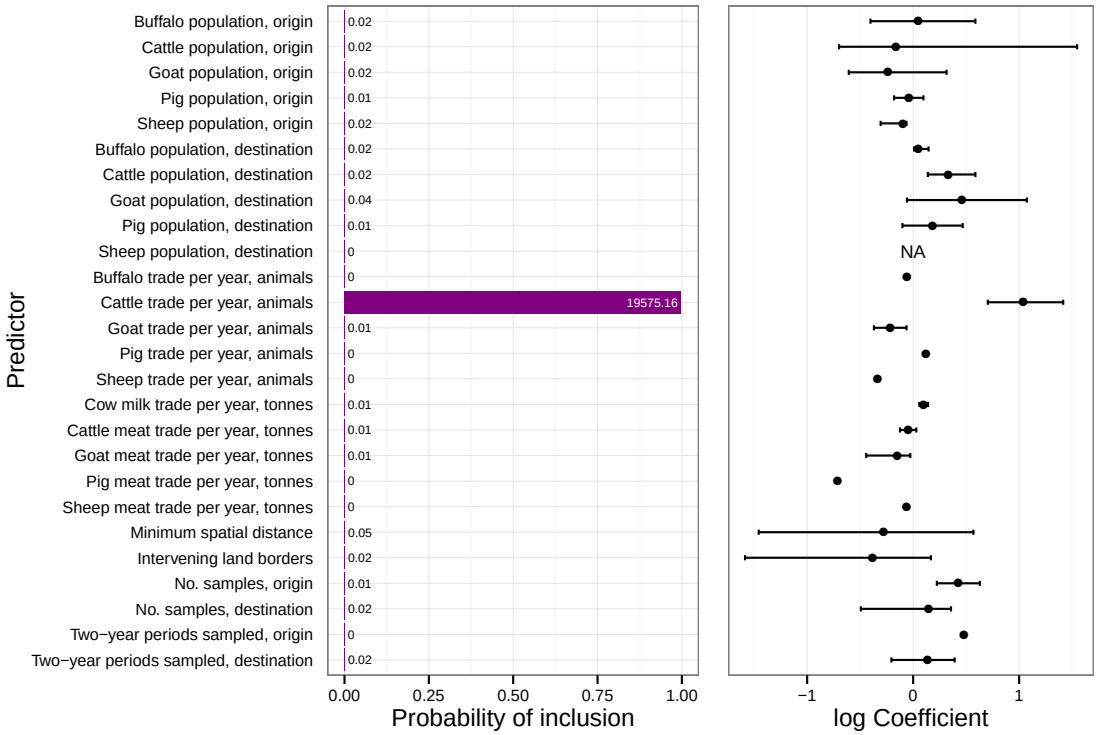


Figure 4.8: Predictors of global topotype SEA diffusion. The graph on the left shows the posterior probability for the inclusion of each predictor in the model; each bar is annotated with the Bayes Factor support for its inclusion. The graph on the right summarises the posterior distribution for the coefficient of each predictor, when that predictor is included in the model. Where NA appears, the posterior probability for inclusion of the predictor was 0.

the geographical Markov Jumps reconstruction; here I present the results for all ten sampling replicates (designated SEA1-SEA10). Arrows are drawn between countries for which there was a posterior probability of at least 90% that at least

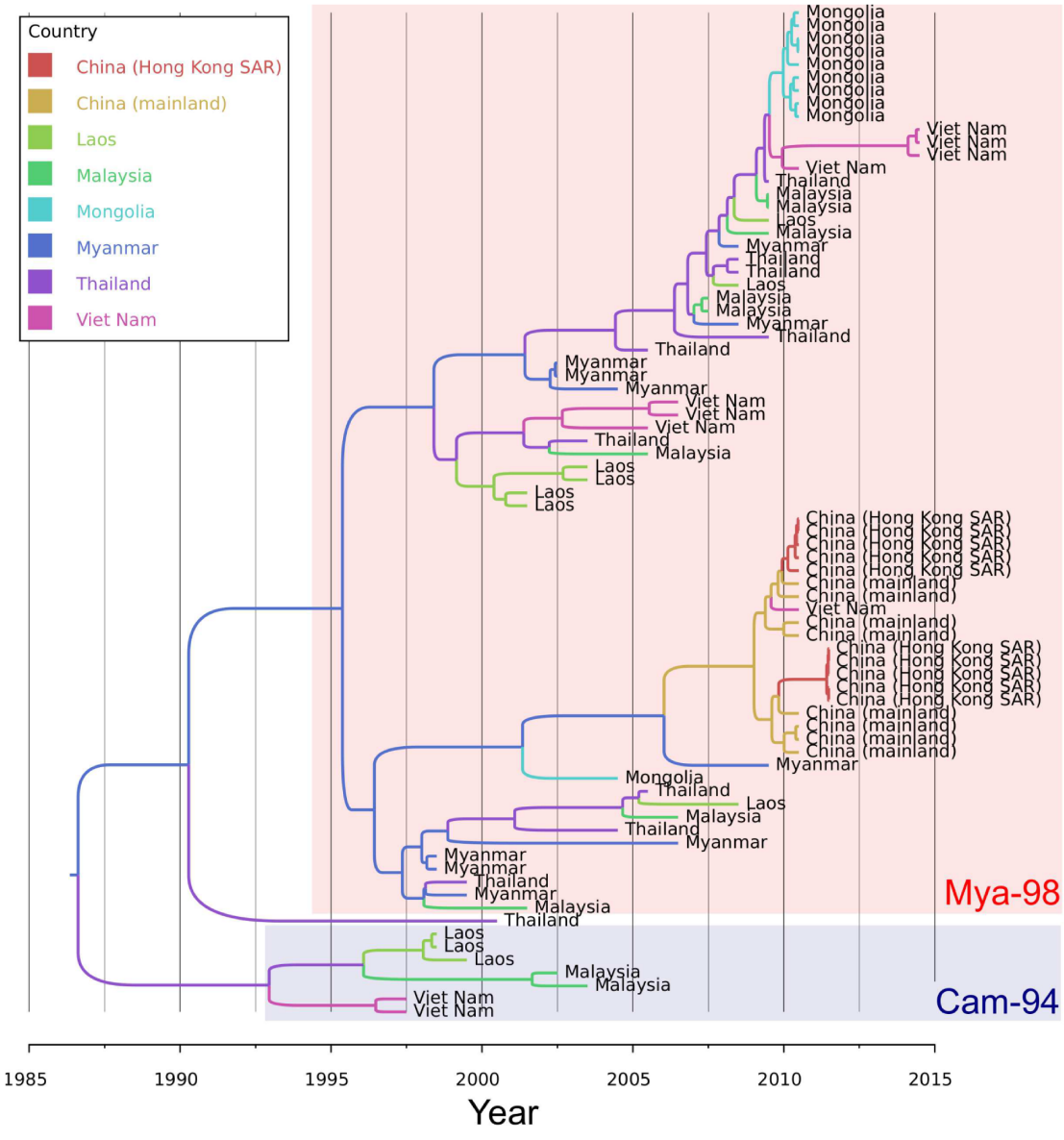


Figure 4.9: Maximum clade credibility phylogeny of an analysis of 78 topotype SEA sequences. Tips are labelled by, and branches are coloured by, the reconstructed country of origin. Clades representing the Mya-98 and Cam-94 lineages previously identified are annotated.

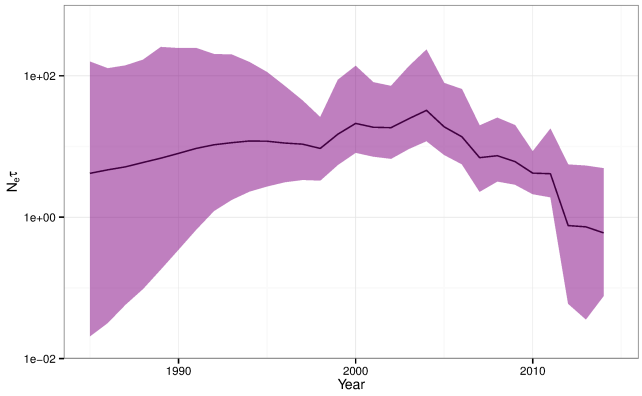


Figure 4.10: Reconstructed skygrid plot for analysis of the SEA topotype. The black line follows variation in the the posterior median population size through time, and the coloured area marks the 95% HPD intervals.

one jump in that direction occurred, and are coloured by the posterior median number of jumps. While the median values vary, some links were inferred in every replicate: from China to Hong Kong, from Myanmar to Thailand, and from Thailand to Myanmar, Laos, Viet Nam and Malaysia.

The NCBI metadata provided the host species for 185 of the 193 SEA sequences with sampling dates between 1995 and 2014. One sequence from a gazelle was excluded. Of the remainder, 123 (66.8%) were from cattle, 48 (26.1%) from pigs, and 13 (7.1%) from *B. bubalis*. A replicate of the sampling scheme produced 73 sequences. Posterior probabilities for the root host species were 0.726 for cattle, 0.201 for pigs and 0.0726 for buffalo. The TMRCA estimate in this case was April 1987 (January 1980-January 1993), and the estimated parameters of the molecular clock were a mean of 7.59×10^{-3} (5.68×10^{-3}-0.0102) and a standard deviation of 5.05×10^{-3} ($2.6 \times 10^{-3} - 8.73 \times 10^{-3}$) substitutions per site per year. These figures do not differ greatly from those from the phylogeography analysis. The MCC phylogeny, with branches coloured by reconstructed host species, can be seen in figure 4.12.

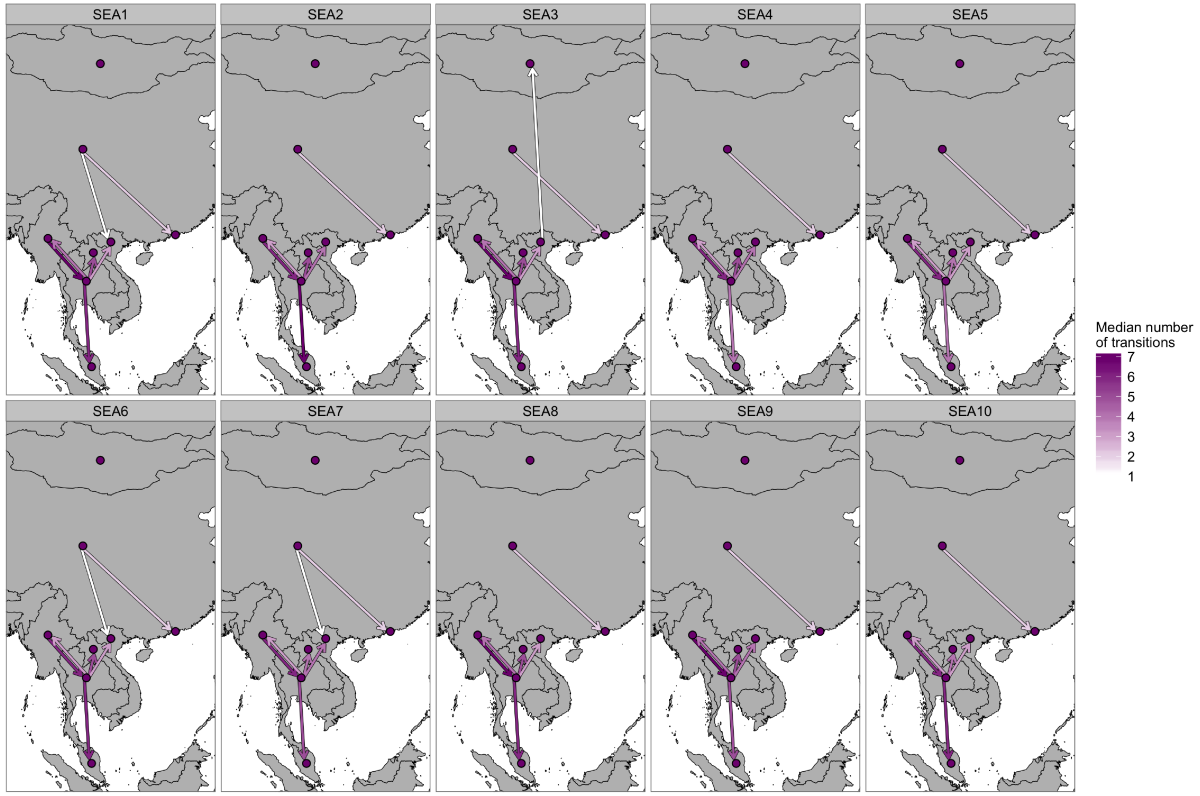


Figure 4.11: Summary of the Markov Jumps reconstruction of geographical movements for the analysis of the SEA topotype. Points indicate countries that provided sequences that were included in the analysis. Arrows are present where there was a posterior probability of at least 90% that at least one reconstructed transition occurred in that direction, and are coloured by the posterior median number of such transitions.

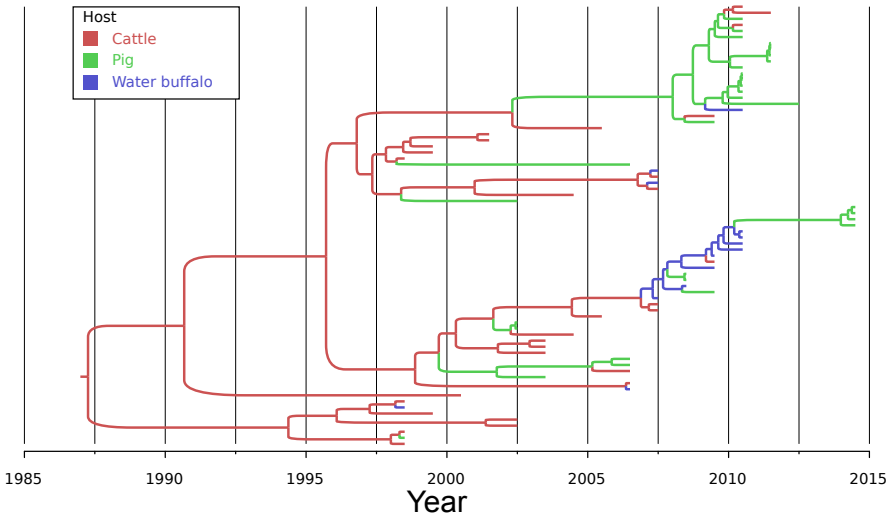


Figure 4.12: Maximum clade credibility tree for 73 topotype SEA sequences. Branches are coloured by reconstructed host species.

The Markov Jumps reconstruction of transitions between host species, once again for all ten sampling replicates (SEAH1-SEAH10) can be seen in figure 4.13. At least one of three types of transition were reconstructed with 95% posterior probability in every replicate: cattle to pigs, cattle to *B. bubalis* and pigs to cattle. One replicate (SEAH6) reconstructed at least one of every single type of transition with 95% posterior probability. Jumps from cattle to pigs were the most common in all but one replicate with median counts ranging from 6 (95% HPD: 2-10) to 9 (7-12).

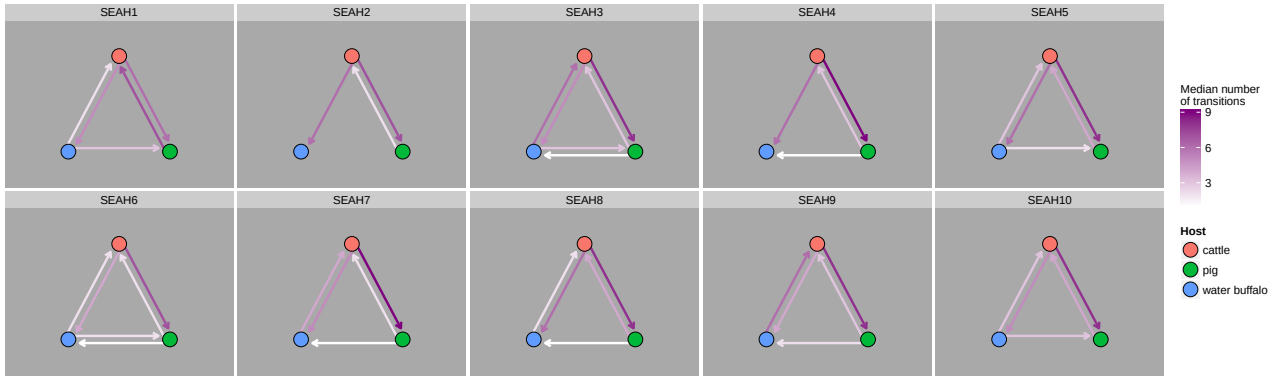


Figure 4.13: Summary of the Markov Jumps reconstruction of host species movements for the analysis of the SEA topotype. Nodes represent species; arrows are present where there was a posterior probability of at least 90% that at least one reconstructed transition occurred in that direction, and are coloured by the posterior median number of such transitions.

4.3.3 Topotype ME-SA

The ME-SA cluster was the largest identified by UPGMA, comprising a total of 817 sequences. Five were excluded as, in the full serotype analysis, they occurred at the end of terminal branches with very slow clock rates, suggesting that they were the descendants of vaccine strains; these were O/1D/Egypt/Alexandria/2013, O/1D/Egypt/Ismaalia/2013 and O/1D/Egypt/EL-Mania/2013, all of which have 98.9% nucleotide similarity to the vaccine strain O_1/Sharquia/EGY/72 [85] as well as O/IRN/1/2007 and O/Ankara/TUR/31/03/02, which had similarities of 99.1% and 98.4% respectively with O_1/Manisa/TUR/69. Other exclusions from the phylogeography analysis were as follows: 39 for being from isolates sampled before 1995, 107 for coming from outbreaks in countries where FMDV is nowhere endemic (Bulgaria, France, Greece, Ireland, Japan, Republic of Korea, Singapore, Taiwan, and United Kingdom), and 29 for coming from countries providing too few sequences in total (Armenia, Bahrain, Georgia, Iraq, Jordan, Kuwait, Lebanon, Mongolia, Oman, Palestinian Autonomous Territories, Qatar, Russia, South Africa, Syria, and Thailand). This left 637, from 19 countries. These are listed in table 4.2; Viet Nam and Laos did not provide detailed trade data and hence trade between them was set to zero in the predictor matrices. The sampling scheme for ME-SA selected 176 per replicate. Table 4.3 displays correlation between predictors. Strong correlations (absolute value of coefficient ≥ 0.7) were as follows:

- cattle population and buffalo population (at either origin or destination)
- sheep population and goat population (at either origin or destination)
- trade in cattle meat and trade in sheep meat
- trade in goat meat and trade in sheep meat
- minimum spatial distance and number of intervening land borders

Predictor	OBP	OCP	OGP	OPP	OSP	DBP	DCP	DGP	DPP	DSP	TB	TC	TG	TP	TS	TCMi	TCMe	TGMe	TPMe	TSMe	MSD	ILB	OSC	DSC	OSP
OCP	**0.79**																								
OGP	0.16	0.45																							
OPP	0.52	0.32	-0.25																						
OSP	-0.13	0.27	**0.72**	-0.39																					
DBP	-0.06	-0.04	-0.01	-0.03	0.01																				
DCP	-0.04	-0.06	-0.02	-0.02	-0.02	**0.79**																			
DGP	-0.01	-0.02	-0.06	0.01	-0.04	0.16	0.45																		
DPP	-0.03	-0.02	0.01	-0.06	0.02	0.52	0.32	-0.25																	
DSP	0.01	-0.02	-0.04	0.02	-0.06	-0.13	0.27	**0.72**	-0.39																
TB	0.1	0.09	0.02	-0.01	0.03	0.04	0.03	0.06	0	0.06															
TC	0.19	0.22	0.14	0.02	0.12	-0.08	-0.08	0.07	-0.1	0.1	0.64														
TG	0.16	0.22	0.18	0.05	0.17	-0.14	-0.14	0.08	-0.18	0.09	0.36	0.6													
TP	0.12	0.13	0.03	0.16	-0.01	0.06	0.05	0.03	0.08	-0.06	0.34	0.32	0.37												
TS	0.09	0.15	0.18	0.01	0.2	-0.19	-0.17	0.1	-0.22	0.14	0.25	0.48	0.7	0.17											
TCMi	0.11	0.13	0.17	0.01	0.19	-0.14	-0.08	0.03	-0.15	0.06	0.36	0.44	0.41	0.35	0.4										
TCMe	0.22	0.32	0.28	0.11	0.24	-0.14	-0.12	0.08	-0.16	0.09	0.21	0.49	0.52	0.1	0.5	0.4									
TGMe	0.15	0.23	0.23	0.03	0.21	-0.2	-0.2	0.06	-0.24	0.1	0.2	0.47	0.51	0.12	0.53	0.37	0.67								
TPMe	0.19	0.3	0.22	0.21	0.15	0.01	-0.02	0.05	-0.01	0	0.07	0.33	0.34	0.23	0.26	0.26	0.66	0.36							
TSMe	0.14	0.29	0.29	0.06	0.29	-0.12	-0.1	0.15	-0.23	0.16	0.19	0.38	0.5	0.07	0.61	0.39	**0.71**	**0.74**	0.45						
MSD	-0.15	-0.23	-0.16	-0.06	-0.1	-0.15	-0.23	-0.16	-0.06	-0.1	-0.27	-0.32	-0.27	-0.33	-0.23	-0.4	-0.25	-0.15	-0.22	-0.2					
ILB	-0.22	-0.32	-0.14	-0.09	-0.03	-0.22	-0.32	-0.14	-0.09	-0.04	-0.17	-0.21	-0.16	-0.24	-0.13	-0.25	-0.15	-0.05	-0.12	-0.08	**0.82**				
OSC	-0.15	-0.1	0.53	-0.34	0.2	0.01	0.01	-0.03	0.02	-0.01	0	0.03	0.03	0	0.01	0.09	0.1	0.05	0.08	0.02	-0.03	-0.02			
DSC	0.01	0.01	-0.03	0.02	-0.01	-0.15	-0.1	0.53	-0.34	0.2	0.01	0.06	0.09	0.04	0.13	0.03	0.05	0.11	0.03	0.12	-0.03	-0.02	-0.06		
OSP	0.14	0.13	0.01	0.12	-0.01	-0.01	-0.01	0	-0.01	0	0.07	0.12	0.1	0.07	0.09	0.19	0.22	0.08	0.14	0.12	-0.06	-0.15	0.34	-0.02	
DSP	-0.01	-0.01	0	-0.01	0	0.14	0.13	0.01	0.12	-0.01	0.01	0.03	0.09	0.03	0.12	-0.07	0.07	0.08	0.04	0.09	-0.06	-0.16	-0.02	0.34	-0.06

Table 4.3: Pairwise Pearson product-moment correlation coefficients for predictors of movement between countries included in the ME-SA analysis. Coefficients with absolute values greater than 0.7 are in bold type. OBP, OCP, OGP, OPP, OSP: buffalo, cattle, goat, pig and sheep populations at origin. DBP, DCP, DGP, DPP, DSP: populations at destination. TB, TC, TG, TP, TS: trade in live animals. TCMi: trade in cow milk. TCMe, TGMe, TPMe, TSMe: trade in cattle, goat, pig and sheep meat. MSD: minimum spatial distance. ILB: intervening land borders. OSC, DSC: sample count from origin and destination. OSP, DSP: 2-year sampling periods represented at origin and destination. Figures are rounded to two decimal places.

In contrast to the situation for SEA, cattle trade was not supported in the GLM analysis (figure 4.14) as a predictor with BF>3, but minimum spatial distance was (BF=83314.81) with a negative relationship to transition rates. No other predictor reached the BF=3 threshold, but both the number of sequences included from the origin country and the number of two-year periods from which a sample was available from that country had BFs greater than 1, supporting their inclusion. (See figure B.15 for the full set of diagrams for each sampling replicate. In contrast to the SEA analysis, in which the predictor results were very consistent across all replicates, there was somewhat more variation for ME-SA. However, spatial distance was always highly supported and any other predictor with BF>3 concerned the nature of the sampling.)

The estimated TMRCA of ME-SA samples from 1995-2014 was March 1988 (April 1982-June 1992). As with SEA, this is not the TMRCA of every NCBI sequence that the UPGMA algorithm assigned to this topotype; some apparently extinct lineages are basal to the clade. (Indeed, while ME-SA is not monophyletic in figure 4.4, all post-1994 sequences are.) Unlike SEA, however, this is not inconsistent with the full serotype analysis, in which the TMRCA of the equivalent clade is January 1986 (October 1980-June 1989). The estimated parameters of the molecular clock were a mean of 6.76×10^{-3} ($5.46 \times 10^{-3} - 8.31 \times 10^{-3}$) and a standard deviation of 5.47×10^{-3} ($3.6 \times 10^{-3} - 7.88 \times 10^{-3}$) substitutions per site per year. It should be noted that the median of a lognormal distribution with these parameters is 5.25×10^{-3}, which is very similar to the estimates assigned to most of the corresponding branches in the ME-SA clade in the full serotype analysis.

Figure 4.15 is the MCC phylogeny. As for SEA, the tree is annotated with notable lineages from the existing literature: PanAsia [85], PanAsia2 [160], PanAsia3 [74], Ind2001 [65] and Iran2001 [85]. Turkey was the most likely root location with a posterior probability of 0.602 and then Iran with 0.125. (The majority of

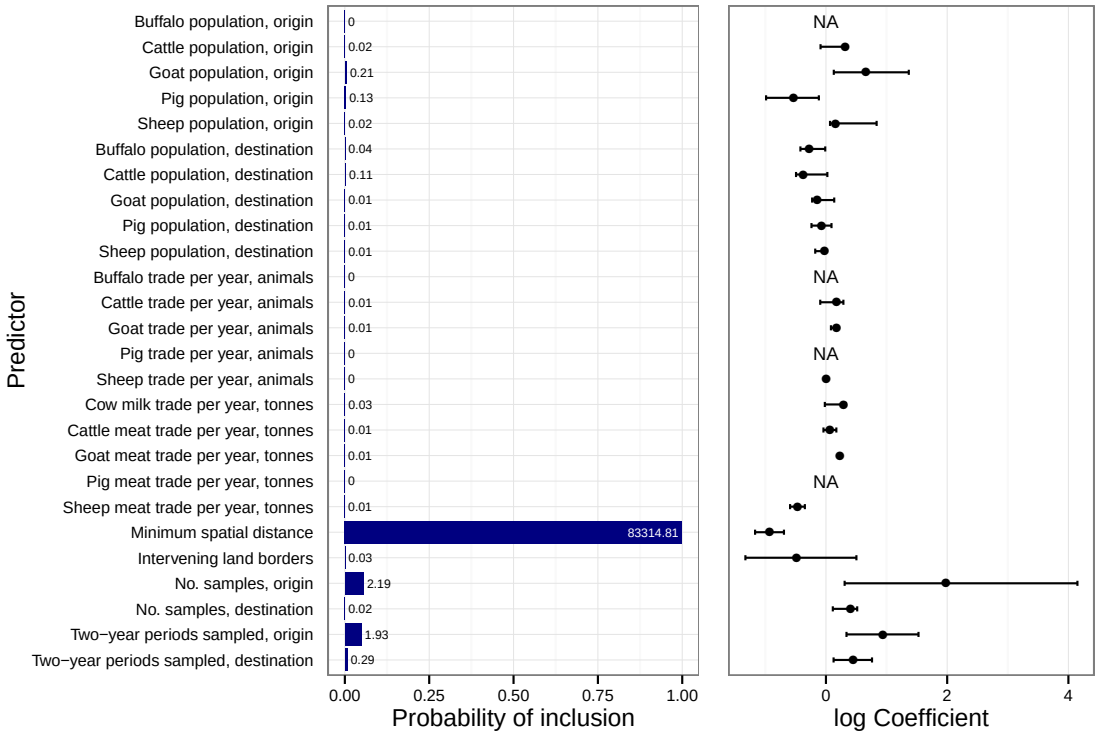


Figure 4.14: Predictors of global topotype ME-SA diffusion. The graph on the left shows the posterior probability for the inclusion of each predictor in the model; each bar is annotated with the Bayes Factor support for its inclusion. The graph on the right summarises the posterior distribution for the coefficient of each predictor, when that predictor is included in the model. Where NA appears, the posterior probability for inclusion of the predictor was 0.

sampling replicates showed similar numbers, but two had a more even distribution of possible locations; see figure B.14.)

Figure 4.16 displays the reconstructed skygrid plots for ME-SA only. As with SEA, a peak in effective population size in the early part of the last decade followed by a decline is observed. Figure 4.17 maps the Markov Jumps as for ME-SA, once again for all ten sampling replicates (MESA1 to MESA10). Here there is considerable variation in which arrows are present between replicates. Notably, however, almost all reconstructed transmissions for which there was 90% posterior support are across a single land border unless there are intervening countries from which no sequences were included in the analysis, the only exception being transitions between Bangladesh, Nepal and Bhutan, and across the Arabian Gulf.

694 of the ME-SA sequences, 1995-2014, had a host identified in NCBI metadata. The five sequences that were likely to have been descended from vaccines were again excluded, as were 13 whose given host species were very under-represented and in some cases vague ("antelope", "gazelle", goat, nilgai, wild boar and yak). Of the remaining 676, 543 (80.3%) were from cattle, 46 (6.8%) from pigs, 30 (4.43%) from sheep and 57 (8.43%) from *B. bubalis*. The sampling scheme selected 120 sequences for each replicate.

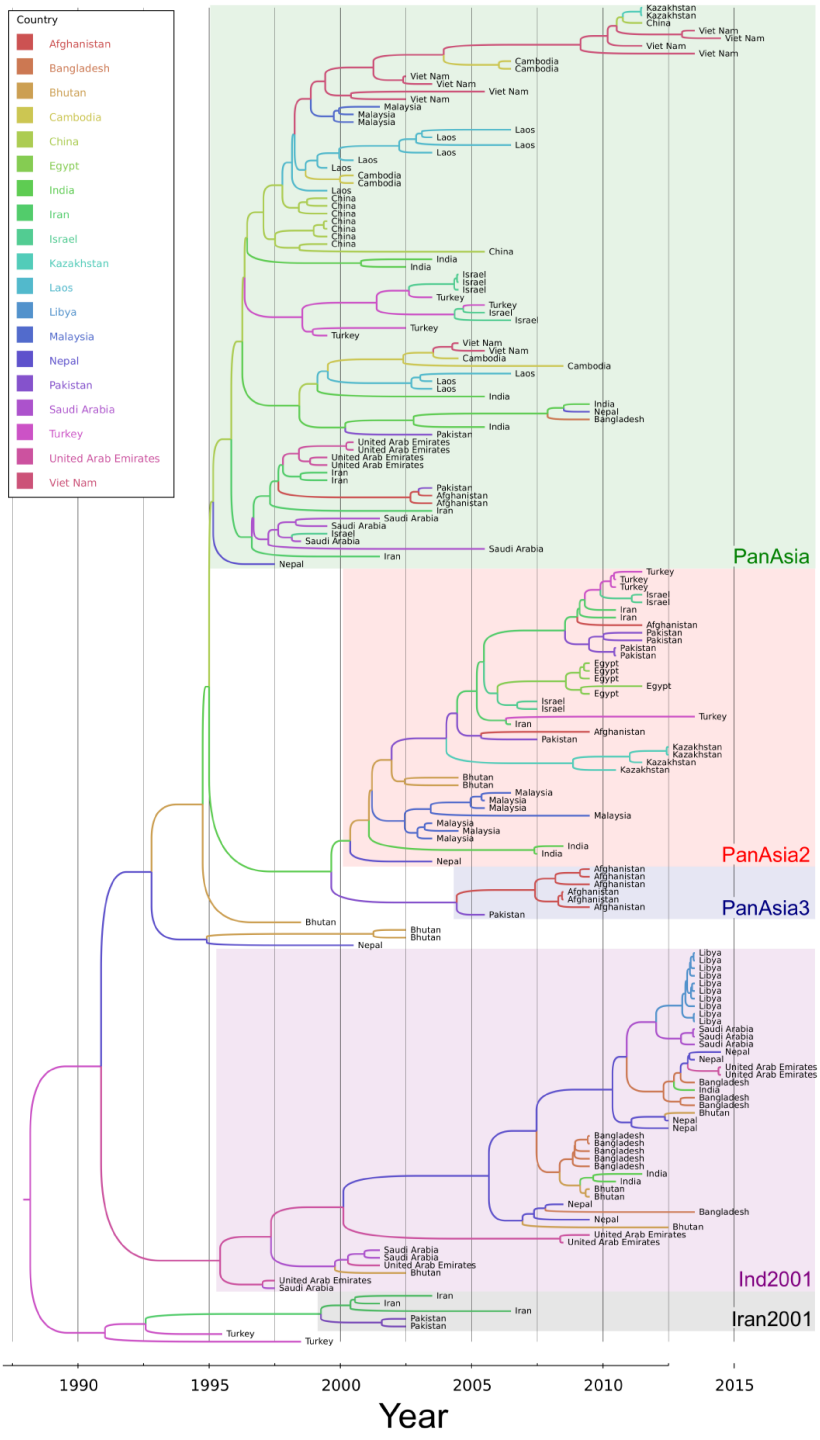


Figure 4.15: Maximum clade credibility phylogeny of an analysis of 176 topotype ME-SA sequences. Tips are labelled by, and branches are coloured by, the reconstructed country of origin. Clades representing lineages previously identified in the literature are annotated.

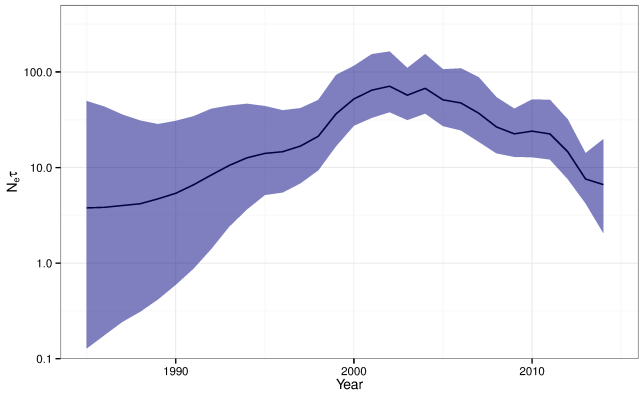
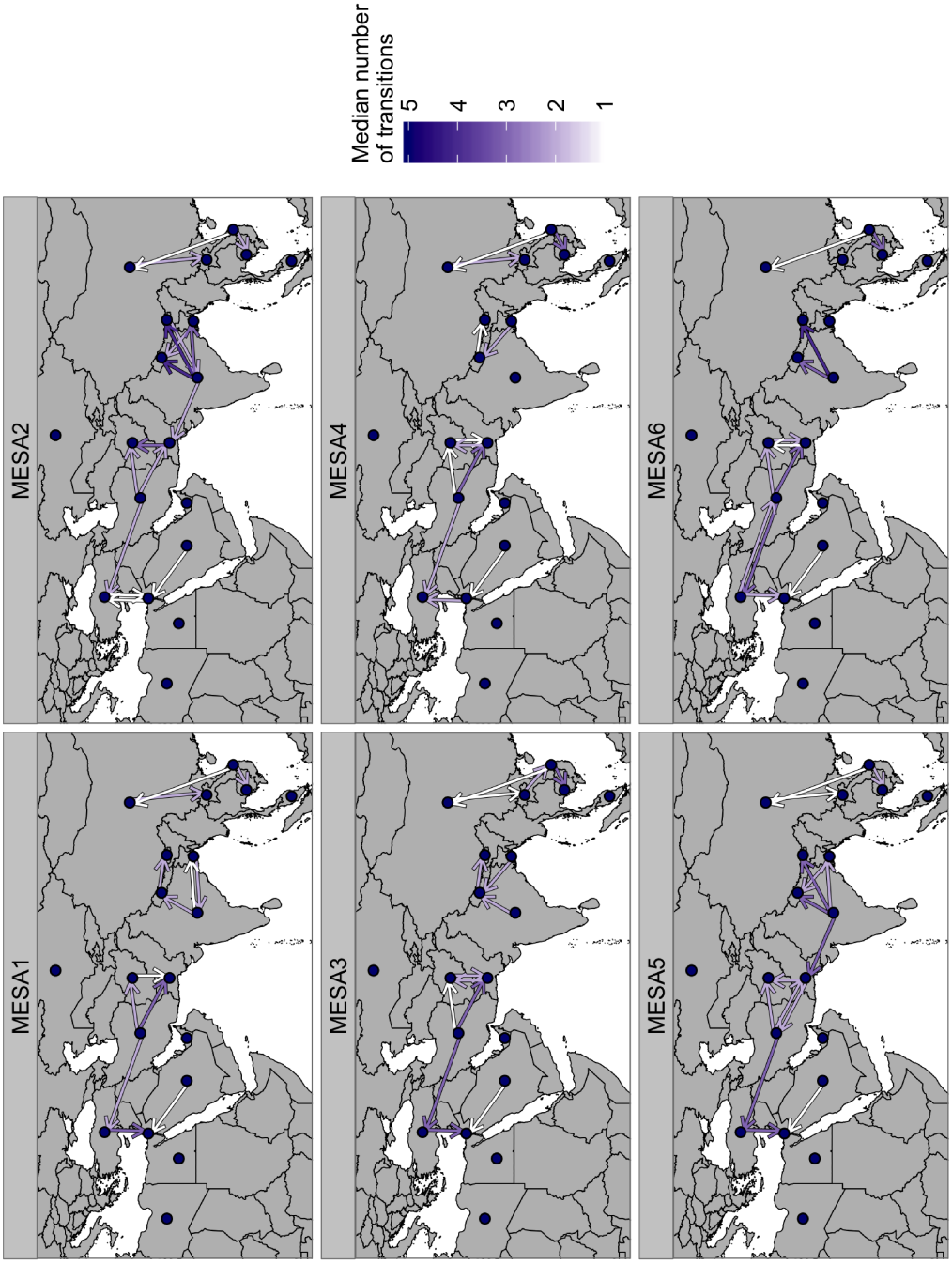


Figure 4.16: Reconstructed skygrid plot for analysis of the ME-SA topotype. The black line follows variation in the the posterior median population size through time, and the coloured area marks the 95% HPD intervals.


Median number
of transitions
5
4
3
2
1
MESA1
MESA2
MESA3
MESA4
MESA5
MESA6

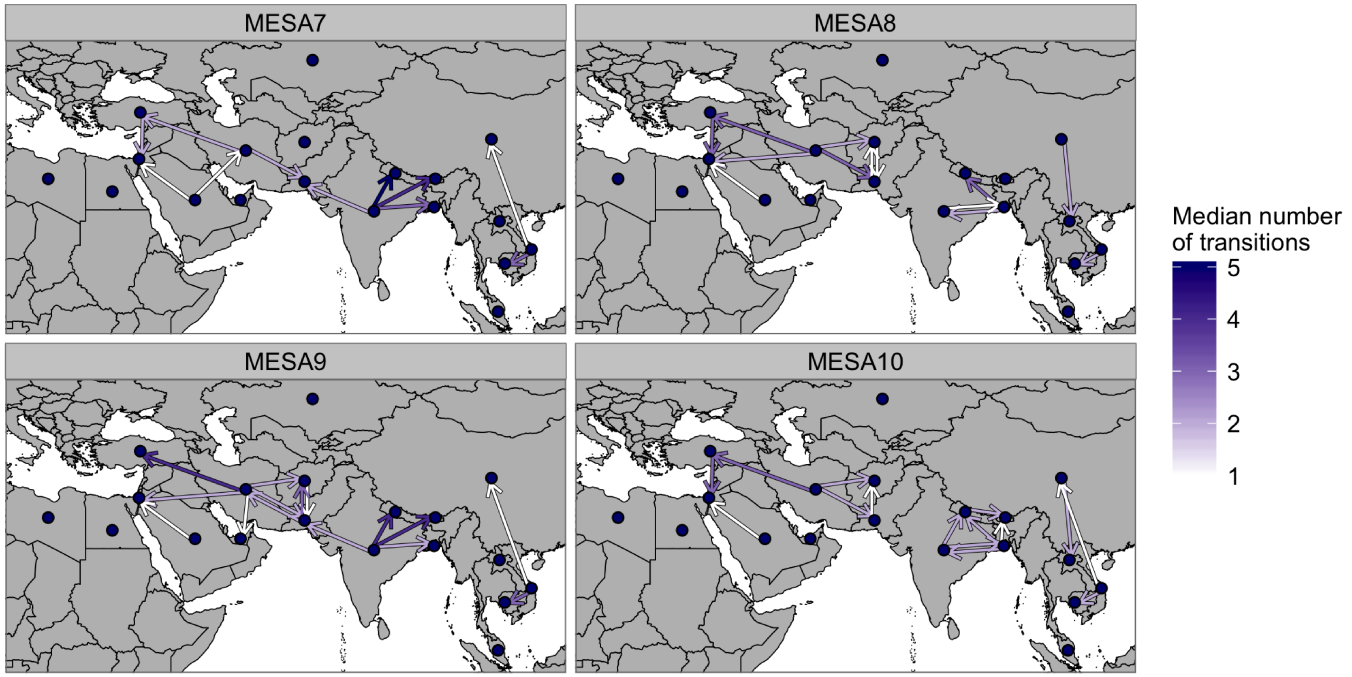


Figure 4.17: Summary of the Markov Jumps reconstruction of geographical movements for the analysis of the ME-SA topotype. Points indicate countries that provided sequences that were included in the analysis. Arrows are present where there was a posterior probability of at least 90% that at least one reconstructed transition occurred in that direction, and are coloured by the posterior median number of such transitions.

The MCC phylogeny from the host analysis can be seen in figure 4.18. The large sheep clade corresponds to the 2001 UK outbreak, sequences from which were very frequently selected as they comprise the bulk of the available sheep data. (Sequences from non-endemic countries were not excluded from the host species analysis.) While this figure might suggest sustained carriage in pigs in part of the tree, this should be interpreted with caution (see Discussion). The most probable root host species was overwhelmingly cattle (posterior probability 0.923). The posterior median root date was January 1991 (February 1987-December 1993), while the estimated clock parameters were a mean of 7.29×10^{-3} ($5.15 \times 10^{-3} - 9.9 \times 10^{-3}$) and a standard deviation of 6.46×10^{-3} ($3.11 \times 10^{-3} - 0.0114$) substitutions per site per year. The TMRCA is somewhat more recent, and the mean clock rate faster, than from the phylogeography analysis but the discrepancy is not large and the HPD intervals show considerable overlap.

Figure 4.19 displays the host-to-host transitions supported by at least 95% posterior probability, once again for all ten sampling replicates. Always supported are transitions from cattle to sheep and cattle to *B. bubalis*, and only one replicate (MESAH2) did not support transitions from cattle to pigs. Transitions from cattle to buffalo were always the most frequent and often considerably so, with median counts from 9 (0-18) to 16 (11-21). While the existence of jumps to cattle, or amongst non-cattle, frequently reached 90% posterior probability, in general these were in much smaller numbers than those from cattle.

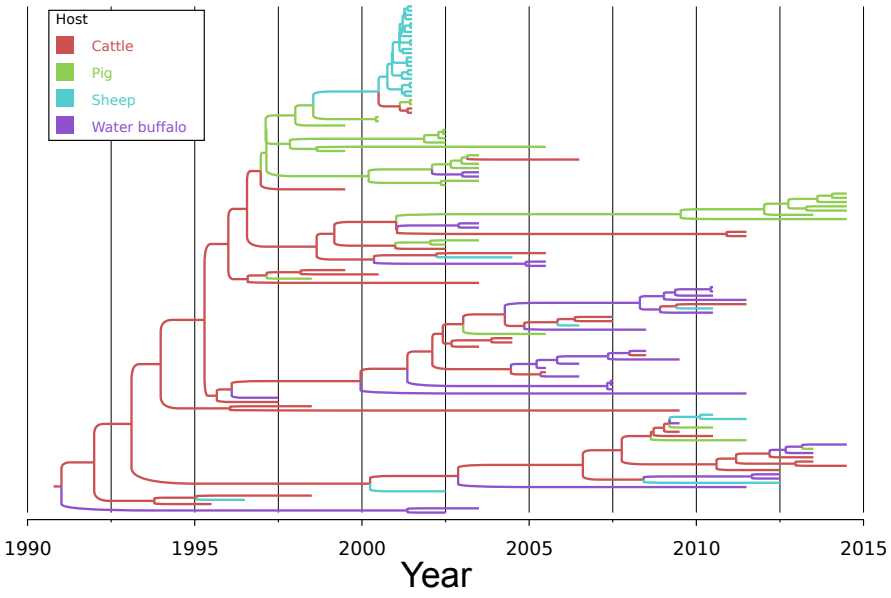


Figure 4.18: Maximum clade credibility tree for 120 topotype ME-SA sequences. Branches are coloured by reconstructed host species.

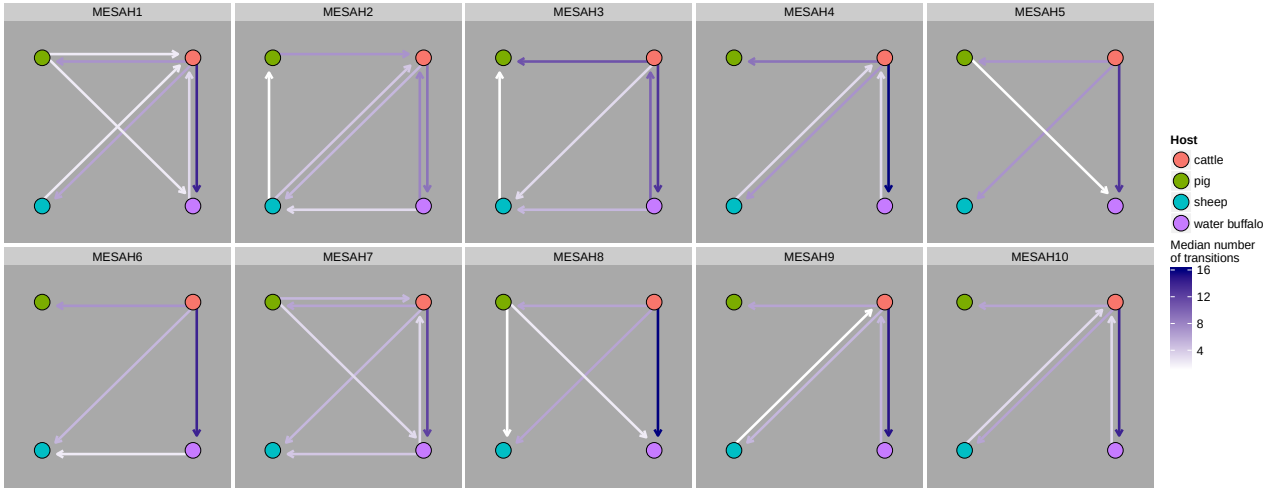


Figure 4.19: Summary of the Markov Jumps reconstruction of host species movements for the analysis of the ME-SA topotype. Nodes represent species; arrows are present where there was a posterior probability of at least 90% that at least one reconstructed transition occurred in that direction, and are coloured by the posterior median number of such transitions.

4.4 Discussion

Unlike FMDV serotypes, which are immunologically distinct and unambiguously defined, the topotypes have only a phylogenetic definition [87, 125]. The UPGMA algorithm allows no ambiguities; every sequence must be assigned to a cluster, but the downside to this approach is that boundaries between clusters are not biologically meaningful. Topotypes, at least in the reanalysis performed here, are in some cases not monophyletic clades in the phylogeny of the entire serotype, and the 85% similarity score for classification is quite an arbitrary choice. (In fact, when the same dataset as used here is classified using a score of 84%, there are only nine "topotypes", with ME-SA, SEA, and large numbers of African sequences forming one and the remaining African data split only into two. On the other hand, when the threshold is 86%, the total number of clusters is 18.) It also appears that the results of applying the same algorithm to modern data results in a different classification from that obtained when the procedure was developed [87], and there is no reason to believe that this will not change again in future. While the utility of these designations to classify currently-existing lineages is not in doubt, a more rigorous definition might be desirable.

That contemporary Euro-SA viruses form two distinct lineages was previously noted by Malirat et al. [99]. The clade that they label "A" is Euro-SA(3) and "B" is Euro-SA(1). In this chapter I showed that a) these are, by the accepted classification, distinct topotypes, b) that there is even a third, extinct topotype amongst the isolates previously assigned to Euro-SA, and c) that the MRCA of these two clusters appears to have been in the very early 20th century. This calls into question how meaningful it is to group them together.

The non-clock-like behaviour of Euro-SA(1) is only likely to be due to the use of vaccines based on O_1 Campos/58 that were improperly inactivated and went on to cause disease. This is also the probable explanation for other tips of the phylogeny

(figure 4.5) which are at the end of terminal branches that are assigned a very slow rate by the clock model; this phenomenon appears to be more widespread worldwide than has been previously reported. Table 4.4 summarises all the isolates for which this ancestry is suspected that did not belong to the Euro-SA(1) cluster, along with the isolate from which they are separated by a branch with a very slow rate, which is likely to be the vaccine strain in question. These results come from all ten replicates of the full serotype sampling scheme, not just the one presented in detail in this chapter. Most of the likely ancestors are indeed known vaccine strains [7, 85], although some (K11/93, O/CAM/6/89 and O Rey Iran/66) are not recorded as being such in the literature. If these genuinely were never used in vaccines, other explanations might be that this ancestor was close to a vaccine strain not included in the analysis, that the virus entered the "dormant" state outside the host proposed by Wright et al. [169], laboratory contamination, or incorrect metadata. As these include Egyptian examples from 2013 [135] and an Iranian isolate from 2007, it appears that vaccines may still be causing outbreaks in some parts of the world.

That some of the contemporary "wild" South American viruses are descendants of the contents of vaccines used decades ago seems likely. This phenomenon violates the normal evolutionary assumptions under which clock-based phylogenetic analysis is performed and should be a consideration for future analyses of this virus. For example, a question is why the distant historical relationship between Euro-SA(1) and Euro-SA(3) has not been reported elsewhere in the literature. De Carvalho et al. [29] performed a molecular clock analysis on all type O sequences for South American isolates sampled from 1994 onwards from South America and found a TMRCA of only 1989. I would suggest that the most likely reason for this is that the existence of the Euro-SA(1) strains descended from vaccines was not taken into account. The set of sequences included in that paper included some which, based on this analysis, are likely to have such an ancestry. On the other hand, no

Isolate	Topotype cluster	Country of origin	Likely vaccine strain
O/1D/Egypt/Alexandria/2013	ME-SA	Egypt	O_1/Sharquia/EGY/72
O/1D/Egypt/Ismaalia/2013	ME-SA	Egypt	O_1/Sharquia/EGY/72
O/1D/Egypt/El-Mania/2013	ME-SA	Egypt	O_1/Sharquia/EGY/72
O/IRN/1/2007	ME-SA	Iran	O_1/Manisa/TUR/69
O/Ankara/TUR/31/03/02	ME-SA	Turkey	O_1/Manisa/TUR/69
O/CAM/1/98	SEA	Cambodia	O/CAM/6/89*
O/CAM/2/98	SEA	Cambodia	O/CAM/6/89*
O/CAM/3/98	SEA	Cambodia	O/CAM/6/89*
O/KEN/10/95	EA-1/3	Kenya	K77/78
K29/95	EA-1/3	Kenya	K77/78
KEN/2/95 (K10/95)	EA-1/3	Kenya	K77/78
K31/07	EA-1/3	Kenya	K77/78
K56/95	EA-1/3	Kenya	K77/78
K121/91	EA-1/3	Kenya	K77/78
K79/02	EA-1/3	Kenya	K11/93*
K51/92	EA-1/3	Kenya	K120/64
O/K/52/1992	EA-1/3	Kenya	K120/64
O_1/N822/Russian Federation/USSR/75	ME-SA	USSR	O Rey Iran/66*
O_1/N850/Russian Federation/USSR/76	ME-SA	USSR	O Rey Iran/66*
O_1/N1491/Russian Federation/USSR/88	ME-SA	USSR	O Rey Iran/66*
O_1/N1467/Jaroslavl/USSR/87	ME-SA	USSR	O Rey Iran/66*
O_1/N1492/Jaroslavl/USSR/88	ME-SA	USSR	O Rey Iran/66*
O_1/N1427/Azerbaijan/USSR/86	ME-SA	USSR	O Rey Iran/66*

Table 4.4: Non-Euro-SA(1) FMDV isolates which seem likely to be the descendants of viruses used in improperly inactivated vaccines, together with the strain suggested to be that vaccine. Those marked with a * are not known vaccine strains.

sequences from the set showing high similarity to O_1 Campos were included, and hence no branches with very slow clock rates would necessarily be detected. The genetic difference between those sequences and any Euro-SA(3) isolate would be smaller than would be expected due to mutation, since in the time between the existence of the Campos strain in the wild in 1958, and its release due to vaccine use many years later, effectively no mutation would have occurred. The result would be that the branches leading to the root would be too short.

The existence of these vaccine-descended viruses has a potentially biasing effect on the estimation of the parameters of uncorrelated molecular clock models. Such a model will assume that branch rates are drawn from a distribution which assigns non-negligible probability to rates that are slower than would be expected in nature, as, effectively, some lineages have gone through long periods undergoing no mutation at all. The inferred parameters of such a distribution would not reflect the variation in mutation rates actually seen in the wild. I recommend an *a priori* check for the possibility that sequences used in an analysis are affected by this and their exclusion. This is less of a problem for a local clock model as slow rates on some branches should not have an undue effect the inference of those elsewhere in the tree. Nevertheless, any molecular clock model is not truly appropriate if the molecular clock is violated, and, here, the fast evolutionary rates inferred for those Euro-SA(1) isolates that may be descended from the vaccines are probably not reliable, as the wide HPD intervals suggest. Any one of countless vaccinations performed over a period of decades could have been the actual origin of these lineages and a model of this sort cannot differentiate between them.

Some of the Euro-SA(1) sequences with high similarity to O_1 Campos are European. The closeness of sequences from the 1967 UK outbreak to that vaccine strain was previously noted by Wright et al. [169]. On the other hand, Danish isolates from 1982-3 have also been reported as close to the European vaccine strain O_1/Kaufbeuren/FRG/66 [25], but this itself has 98.3% similarity on VP1 to O_1

Campos. This rather suggests that the latter vaccine was produced using an isolate derived from an outbreak that was itself ultimately caused by a vaccine. The source of these outbreaks is thought to have been infected meat from South America [169]; it would seem likely that this meat was from animals that were inadvertently infected by the agricultural industry, rather than by naturally-occurring viruses.

FMDV was initially introduced into South America from Europe in the latter half of the 19th century (Samuel and Knowles [125] give 1870 as a date) and both Euro-SA(1) and (3) appear to have been the result of this. Euro-SA(2), on the other hand, consisted of viral populations remaining in Europe, and was also introduced to Mexico. As the topology of the deep branches of the Euro-SA clusters is not particularly well resolved, I draw no firm conclusions about whether (1) and (3) are the result of one or two introductions across the Atlantic. The 1870 date suggests that the TMRCAs for deep nodes in this analysis may be somewhat underestimated. It places the ancestors of contemporary South American strains in the New World before the upper limit of the 95% HPD for the TMRCA of the whole tree, and it seems more likely that Europe was the ultimate source of the entire extant serotype; if it was instead South America then Euro-SA(2) was a reimportation to Europe and there is absolutely no remaining trace of the European viral population that initially seeded South America. It does, however, remain possible that multiple lineages crossed the Atlantic over the course of decades and that earlier introductions died out. The 1870 date could also be for the introduction of serotype A.

A previous study of the Cathay topotype by Di Nardo et al. [34] noted that its molecular clock seemed to be at the fast end of rates estimated for FMDV. The estimate from that paper for the entire topotype, which used an uncorrelated lognormal clock model, was a mean of 0.0106 and a standard deviation (if translated to the real scale) of 0.0102. While this mean would appear considerably faster than the posterior median rate of 8.17×10^{-3} that occurred on most Cathay terminal

branches in this analysis, it should be noted that the *median* of a lognormal distribution with the di Nardo parameters is 7.64×10^{-3}. As was observed with ME-SA, it would seem that the median of the distribution from an uncorrelated lognormal clock model corresponds better with results from a RLMC model than the mean does. This is the first work to note fast rates for this topotype in a framework that simultaneously estimates rates for other lineages, and provides firm support that it is a genuinely fast-evolving lineage. As the Cathay strain is adapted to porcine, rather than bovine, hosts, it is likely subject to different evolutionary pressures from the host immune system, which is the most probable explanation for the difference.

That estimated mutation rates tend to be negatively correlated with evolutionary timescales is well-established, and it is thought that this is the result of both purifying selection and mutation saturation [41, 66]; the former means that mutations are more likely to be observed if a short time has elapsed since they occurred, and the latter means that even neutral mutations will have a tendency to be unobserved as time goes on. This is the probable reason for the discrepancy seen between the TMRCA and rate estimates for SEA between the full serotype and single-topotype analyses. Interestingly, however, this was *not* seen for ME-SA, for which, in the full serotype analysis, the RLMC model was able to determine a rate change point that allowed the results of the two analyses to match closely. (This occurred in only eight out of ten sampling replicates; in the other two a similar situation occurs to SEA.) Likewise, the rate change found for Cathay brings rate estimates here into line with those from Di Nardo et al. [34], and the fact that older Cathay lineages do not show the fast rates seen at the tips mirrors the ME-SA situation. (Local clock models have, I note, previously been suggested as a possible solution to the correlation between rates and timescales [41].) For SEA, and also the African topotypes and Euro-SA(2) and (3), however, the model did not find substantial evidence for a change and those parts of the

tree were inferred using a single rate over the bulk of the timeline. (In fact, a single sampling replicate did indeed find such a break for SEA; see figure B.3j). This likely means that rates at the tips are underestimated (and those deeper in the tree overestimated) and that the results of the individual SEA analysis are more reliable.

The RLMC model as implemented in BEAST [40] has been rather neglected in phylogenetics research, but this chapter shows it has obvious utility. The assumption that clock rates are uncorrelated over the entire tree fails in general because faster rates are expected over shorter timescales, and also fails if mutation rates would be expected to vary due to ecological factors (such as host species). The reason for the lack of attention may be that its current implementation displays poor mixing behaviour. I establish here that this can be improved by the use of Metropolis-coupled MCMC, but, even so, the size of the dataset used was considerably smaller than could be accommodated in an analysis with an uncorrelated clock model; had I used the latter, I could have at least doubled the size of the datasets and achieved convergence in reasonable time. There is, then, scope for further methodological work, improving MCMC tree proposals for exploring the space of phylogenies with branches whose assigned clock rates are correlated with those assigned to their neighbours.

When using publicly-available data, as I did in this chapter, identification of problematic sequences *a priori* is a potentially laborious task. None of the five examples that were excluded from the ME-SA analysis had any information in their NCBI metadata that suggested their likely descent from a vaccine strain. While the Egyptian sequence metadata did refer to a paper in which their similarity to O_1/Sharquia/EGY/72 was discussed [135], the other two did not. Although these five sequences are slight outliers on the root-to-tip divergence plot for ME-SA (figure 4.20), this effect is not dramatic and this was not enough to prompt their automatic exclusion. With exhaustive checking of every sequence used in

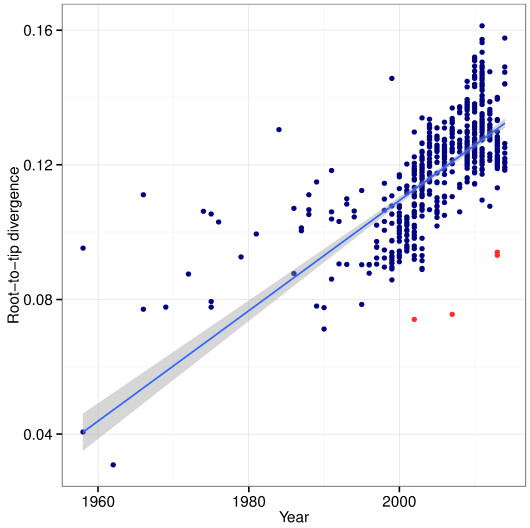


Figure 4.20: The ME-SA graph from figure 4.3 with points representing the five sequences that were probably part of outbreaks resulting from vaccine strains highlighted in red.

an analysis likely to become increasingly time-consuming for researchers as the sheer amount of genetic data increases, and with the great deal of computational time involved in running a modern phylogenetic analysis on a large dataset, an argument can be made for a collaborative effort amongst researchers to update and maintain detailed metadata, in order to prevent duplication of effort.

Some of the difficulties in constructing a consistent sampling scheme using retrospective data from publicly available databases are illustrated here. In the previous chapter I argue that it appears to be preferable to stratify sequences by both location and year, but strictly applying this proved impossible, and the procedures outlined here are inevitably a compromise. For the full serotype analysis, I wished to prioritise the use of as much historical data as possible, but data prior to 1995 is so sparse that an attempt to stratify by any variable is incoherent. The scheme in that case was merely intended to reduce the possibility that the spurious bottleneck effect identified in chapter 3, in which multiple sequences taken from the same area at the same time bias coalescent reconstructions, was present.

For the phylogeography analyses I chose instead to exclude older sequences, and some degree of stratification became possible. While chapter 3 suggests that equal groups are preferable, a strict insistence on a minimum of ten sequences per country would have excluded potentially crucial locations from the analysis (notably China), so I included those for which at least five were available. For many countries, the NCBI database included sequences for a wide variety of sampling dates, but this was not always true; examples are China, Hong Kong and Mongolia for SEA, and China and Libya for ME-SA, for which most or all sequences were sampled in a single year and are very closely related, forming a single clade. I observe that few links involving these countries were reconstructed using Markov Jumps. When all sequences from one location form such a clade, only one jump into that location is necessary to explain the data, and in most cases it seems that the origin of that jump was uncertain. Ideally, samples widely dispersed in time would be available from all countries of interest.

Another consequence of the sampling scheme is that, when ten or less sequences per location (or thirty or less per species) were available, each replicate of the sampling scheme would select them all. The variation in the contents of the sequences sets between sampling replicates is much greater for well-sampled countries, and it is possible that the some degree of the general consistency in parameter estimates and reconstructions seen between replicates (see appendix B) is the result of data that was included in every single one. Nevertheless, it is reassuring that the estimates of numerical parameters and the skygrid reconstructions did not differ greatly between replicates, especially as, due to the quantity of data available and the sampling schemes selected, I was forced to use fewer sequences than the 200 recommended in chapter 3.

The behaviour of the SEA topotype in south-east Asia that is reconstructed here matches the scientific consensus on the epidemiology on this strain closely. Cattle in Myanmar appear to be the reservoir, and the virus is spread by some

of this population being moved for trade into Thailand and beyond [1, 33]. The cattle trade is the engine that drives viral movement and is therefore the only predictor required to explain it. I reconstructed regular transitions from Myanmar to Thailand (a posterior median of six or seven in every replicate), and smaller numbers from Thailand to Malaysia, Viet Nam, Laos, and back to Myanmar. (There are only 3 Cambodian sequences available for SEA in the NCBI database for the time frame here, all from 1998, and thus the country did not meet the threshold for inclusion.) While Thailand does import cattle from its other neighbours [33], this does not appear to be important. It should also be noted that Myanmar, Laos and Viet Nam are not countries that report trade data for FAOSTAT and hence the values in the trade predictor matrix for trade between them were set to zero, but the GLM model still did not require anything additional to this matrix to explain viral movement; therefore, it is reasonable to infer that these six country-to-country transitions are not relevant to the movement of infected animals. The reconstruction in the rest of the SEA area is less detailed as a result of the paucity of Chinese data; there are only eight eligible Chinese sequences for SEA in the database, all are from 2010, and all are closely related. While the transmission of mainland Chinese viruses to Hong Kong is strongly supported and to be expected given a considerable volume of trade in this direction, it is unclear whether the country is an important source of infections anywhere else. The link suggested from Viet Nam to Mongolia in replicate SEA3 seems unlikely to be direct and is probably the result of Chinese diversity that is missing from the analysis.

The fact that trade was not found to be a well-supported predictor of lineage movement for ME-SA, unlike SEA, is somewhat surprising. As only Laos and Viet Nam are not reporters to FAOSTAT from that analysis, responsible for only two entries in each predictor matrix, it seems unlikely that this is the reason. It may be that the FAOSTAT data, whose quality is not faultless, does not adequately

capture formal trade patterns in that part of the world, and that as a result spatial distance was a more reliable predictor. It may also be that some predictors are important in some parts of the affected area but not in others, and that this model is not well-equipped to separate these; for example, if SEA is spread in its affected area by cattle trade it seems unlikely that ME-SA, which is present, is not also, but in the Middle East it may be that trade in small ruminants is more significant. The GLM model would need to be refined to be able to handle a situation of this sort. Finally, it may be that animal movements between some of these countries are more casual and thus unlikely to be picked up in official statistics at all. For example, border control between Iran, Afghanistan and Pakistan is weak and many herds are nomadic [131]; the existence of a continuous population of small ruminants in these areas and west to the Mediterranean basin has been dubbed "Ruminant Street" [33], a phenomenon which has been hypothesised to drive FMDV viruses originating in south-central Asia in a westerly direction [122]. In the reconstructions here, however, an entirely westward pathway is only reconstructed in two replicates (MESA5 and MESA9). On the other hand, a key role for Iran in spreading viruses both westwards and eastwards is almost always evident.

A question that arises concerns the status of ME-SA in the areas where SEA is also present. The dearth of Myanmar sequences for ME-SA in the NCBI database precludes investigation of the role of that country in transmission of the topotype, but only one ME-SA isolate (from 1999) has ever been discovered there (NJ Knowles, personal communication), suggesting it is rare. SEA is also much more regularly recorded than ME-SA in Thailand. Viral populations of ME-SA in south-east Asia are therefore more likely to be maintained elsewhere, in Laos, Viet Nam or Cambodia [122], and may have arrived from China in the late 1990s [85]; its failure to establish itself in Myanmar and Thailand could be due to smaller trade volumes in those directions, or potentially strain competition. As once again

few Chinese sequences are available, conclusions about the locations of ME-SA populations in the region must be tentative, but transitions from Viet Nam to Cambodia are always well-supported, and transitions from China to Laos usually so. There does not appear to be much mixing between southern and south-eastern Asian lineages, perhaps because the Himalayas and the Arakan Range provide natural barriers, although Rweyemamu et al. [122] report that some volume of trade from India and Bangladesh into Myanmar does occur.

It has been noted that some ME-SA strains in Malaysia belong to the PanAsia2 lineage, which is not otherwise encountered in South-East Asia [1], and that this may have been due to an import of beef or buffalo meat from India; tips from this lineage can be seen in figure 4.18. This analysis did not reconstruct this event at the 90% threshold in any replicate, but it was often quite probable; posterior probabilities for at least one jump from India to Malaysia were over 0.5 in six out of ten replicates. That trade in cattle meat (statistics for buffalo meat are not available in FAOSTAT) was nevertheless not selected as a predictor is likely due to the fact that transmission due to contaminated meat is an unusual route of infection, not one occurring with regularity over the virus' entire range.

Elsewhere, routes of introduction for ME-SA to the Arabian peninsula are not identified with clarity in this analysis. Bangladesh, India, China, Iran and Turkey are all frequently-reconstructed origins for jumps to Saudi Arabia and UAE, but it is rare that any obtains 90% support, and variation between replicates is considerable. Mixing of lineages between India, Nepal, Bhutan and Bangladesh appears to be complex. One import to Israel seems clearly to have come from the Arabian peninsula, as previously reported [141], while others are generally traced back to Turkey, although, with no sequences included from the country's other neighbours, the exact route is unclear.

While it is very plausible that Myanmar was the ultimate origin of the SEA strain,

and there is no obvious reason that the data here would be biased in favour of that hypothesis, the strong suggestion that the common ancestor of all ME-SA isolates sequenced in the last twenty years was located in Turkey should be treated with caution. This is because the only ME-SA sequences for isolates collected in 1995, the earliest year eligible for inclusion, were Turkish; this likely introduces some bias. (In contrast, the earliest year for which SEA sequences were available was 1997, all from Viet Nam, and 1998 sequences are available from every other country in south-east Asia.) As this preference was in fact only shown in eight out of ten sampling replicates, with the other two showing considerable uncertainty regarding the root location (see figure B.14), and with the accepted placement of Turkey at the end of "Ruminant Street", receiving viruses originating further east, this finding would need better support than is demonstrated here. Given the nature of the available data, it is likely impossible to design a sampling scheme to minimise bias in answering every potential question of interest.

There is a mismatch in the GLM analysis between the time period that it focussed on and the parameters that it infers. Samples were included from the last 20 years, and the data used to inform predictors was also from this period. The model takes the geographical movement rates, however, to be constant over the entire history of the phylogeny. As datasets become larger and models more complicated, concerns like this may become more prevalent. A possible solution in this case would be to develop a model that combined GLM predictors with the epoch model [13], which allows separate transition rate matrices for pre-defined time periods; the rates and reconstruction for the 20 years of interest would be the desired output and those in the earlier period, which will not be informed by a large amount of data, would be discounted. This method could also be used to subdivide the timeline even further, with different predictor matrices calculated from different data for each period.

All the skygrid reconstructions suggest that FMDV populations have been in

decline for around a decade; the HPD region in each is unambiguous and would not accommodate a straight line. While the spurious bottleneck effect described in chapter 3 cannot be entirely ruled out for the topotype-specific analyses, as some locations provided many closely-related samples from the same year, the care with which the sampling schemes were constructed at least makes it less likely. The topotype-specific reconstructions also show a more gentle decline than the sudden drop-off seen in chapter 2. If these declines are indeed genuine, it suggests that agricultural development in the regions affected may have been successful in decreasing incidence; no case report-based epidemiological studies of global FMDV incidence over this period appear to have been published, and a comparison may be instructive. The gradual disappearance of FMDV from South America over the past decades would be expected to have an effect in the reconstruction of the full serotype. While this plot suggests two clear peaks, I showed in chapter 3 that such reconstructions can be misleading; the HPD intervals are also consistent with a constant population size for the last quarter of the 20th century. The peak observed in the ME-SA plot sometime between 2000 and 2005, however, is seen even in the HPD region. This may be connected with the rapid spread of the PanAsia strain [85] around this time.

The 90% posterior probability cut-off for the identification of well-supported transitions between countries is quite strict, and certainly favours specificity over sensitivity in identifying links; there was still considerable variation between replicates for ME-SA in particular. Because specificity is favoured, the absence of an arrow from the map should not be taken to suggest that transitions in that direction have not occurred. Nevertheless, this variability is a cause for concern, as the exact set of sequences used clearly makes a difference, and few published phylogeography analyses test their results for robustness to the particular set of included sequences. I would suggest that the era during which this might have been acceptable should be coming to an end. Choosing every single sequence

that is available for a particular virus will become increasingly expensive in terms of computational time, and in any case, the resulting dataset will have been constructed with no regard to stratification by any variable of interest. If downsampling is performed in order to deal with this problem, a single replicate of the procedure should not be regarded as providing definitive results. Another feature of these reconstructions is the occasional implausible inference, such as the direct link from Viet Nam to Mongolia in SEA3, and from Saudi Arabia to Iran in MESA7 (contrary to the direction of trade). Results like these would best be confirmed in multiple analyses before they can be accepted as likely to be genuine.

In both topotype analyses, the dominant GLM predictor shows a strong correlation with another predictor whose inclusion was nevertheless not supported: for SEA trade in cattle correlates with trade in goats (correlation coefficient 0.71) and for ME-SA minimum spatial distance correlates strongly with number of intervening land borders (0.82). (No member of any other pair of strongly correlated predictors was ever well-supported for inclusion.) This suggests an encouraging degree of robustness in the GLM method to collinearity between predictors. For SEA in particular, any identified relationship between goat trade and FMDV diffusion would have to be treated with caution as this correlation makes cattle trade an obvious potential confounder, and this model, indeed, appears to control for it correctly, with goat trade receiving no more support than any of the other excluded predictors.

The host jump reconstruction shows similar between-replicate variability to the phylogeographical reconstruction, but the overall pattern is inescapable: cattle are the most important reservoir for viruses of both topotypes, regularly transmitting to other species, and while movements between others species and back to cattle do occur, they are much rarer. The MRCA of all isolates, 1995-2014, from both clusters, was also almost certainly infecting a cow. While the great majority of samples were isolated from cattle, this is not the reason for these findings as I

included equal numbers from other species (except in the case of *B. bubalis* for SEA, as there were not enough available to do that). The host species analysis does illustrate some problems concerning sampling based on incomplete historical data. A particular example can be seen in the MCC tree for ME-SA (figure 4.18), which suggests sustained carriage in pigs in part of the phylogeny. This is misleading. The reason that this occurs is that, in attempting to provide equal numbers of isolates from each host for the analysis, the algorithm picked large numbers of pig sequences from south-east Asia, simply because there were more of these available due to the much larger pig populations in that region. Cattle sequences, on the other hand, are more widely distributed over the entire range of the topotype. As a result, the East Asian portion of a random sample of sequences from the topotype's entire distribution will be disproportionately from pigs. For comparison, figure 4.21 is from an analysis of only south east Asian sequences, with equal numbers included from cattle and pigs; while it suggests sustained transmission in pigs is possible in some parts of the tree, there is no suggestion that pigs are primarily responsible for maintaining the viral population. A solution to this would be to stratify by location and by species (and also by time, at least as much as possible), but if this is attempted in this dataset, sequence numbers become very low indeed. A separate oddity with this sampling scheme regarding ME-SA is that in attempting to provide equal numbers of sheep sequences, the algorithm picked large numbers from a single outbreak, the 2001 British emergency. It is likely that this is no better or worse than including a single sequence from this outbreak and having greatly unbalanced groups, but as in chapter 3 it was found preferable to use equal group sizes and not attempt to weight numbers by any measure of population size, neither is ideal.

In summary, what I aimed to present here was a phylodynamic analysis of serotype O using a methodology appropriate for the era of large datasets; no longer is it possible or appropriate simply to use everything that is available. Some of the

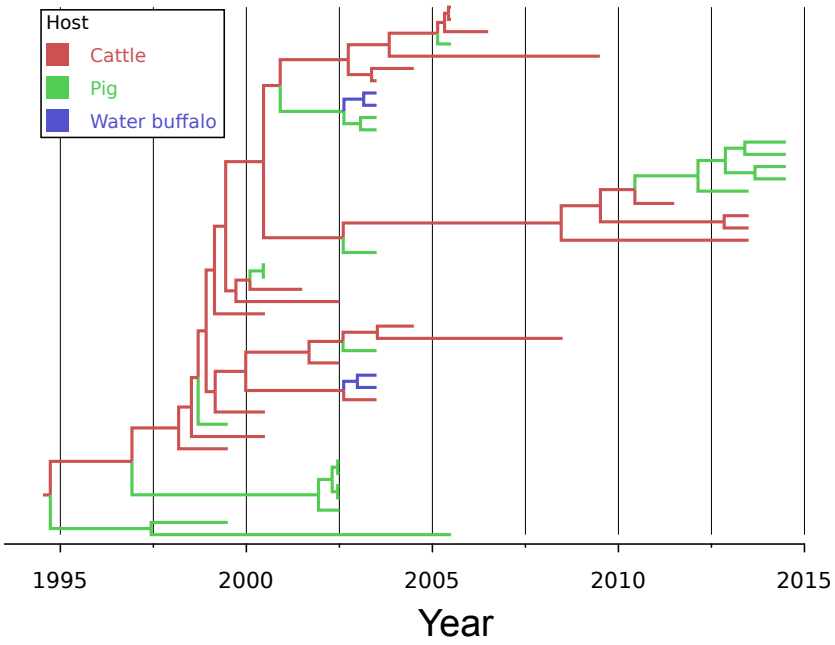


Figure 4.21: Maximum clade credibility tree from an additional analysis of 44 topotype ME-SA sequences from south-east Asia, chosen such that 20 sequences were included from both cattle and pigs. Branches are coloured by reconstructed host species.

issues arising from instead choosing a subsample are illustrated. Firstly, different selections of sequences randomly chosen under the same methodology can lead to quite different reconstructions, particularly when counting transitions. Secondly, any researcher examining historical patterns is constrained by the rather haphazard way that isolates were selected for sequencing in the past. Ideally, in the future, isolates will be collected more methodically. The results of the analysis showed that it seems likely that there are two distinct lineages present in South America at the current time, and that at least some of the members of one of them are the descendants of viruses from improperly inactivated vaccines in use in the latter half of the 20th century. Previous suggestions that the pig-adapted Cathay topotype is particularly fast-evolving are confirmed, as is the importance of Myanmar in maintaining populations of the SEA topotype. It also seems that serotype O populations are in decline both globally and in the regions affected by both ME-SA and SEA.

Chapter 5

Simultaneous exploration of the space of phylogenies and transmission trees: theory

5.1 Introduction

The remaining chapters of this thesis move from one end of the phylodynamics scale to another, from the exploration of global patterns of dispersal to the reconstruction of transmission trees from a single epidemic. The various extant methods available for this purpose were described in chapter 1. I contribute a new method of the third type described there, accommodating both within-host mutation and within-host diversity. I follow Cottam et al. [26] and Didelot et al. [35] in finding transmission trees by annotating the internal nodes of a phylogeny (see figure 1.3). Those two studies were constrained by the lack of a method to co-estimate the complete phylogeny simultaneously with its node labels; they instead employed a "two-step" procedure, using a fixed tree pre-generated by a standard phylogenetic

method. This approach has two problems. Firstly, it will ignore any uncertainty in estimates of the phylogeny. If a Bayesian phylogeny reconstruction method is used, this can be mitigated by using the same method on each one of a sample of trees drawn from the posterior distribution, but at the cost of greater computational time. Secondly, the method used to construct such a fixed tree will often have made assumptions about the structure of population of pathogens or infected hosts that are inconsistent with that of an epidemic. Standard analyses for estimation of time-resolved phylogenies, such as the skyline family used in chapters 2 to 4, assume that all lineages are part of a single, freely mixing population, with the probability of a tree calculated based on the assumption that it was generated by a coalescent process in this population. The result is that phylogenies may display features that are not epidemiologically plausible. For example, even for the fastest-evolving RNA viruses it remains true that many sequences collected over the short timescale of an epidemic will be identical [10]. If this is the case for two isolates, they are likely to form a "cherry" in the reconstructed phylogeny whose TMRCA can take values very close to the sampling time of the earlier isolate, because in a panmictic population, there is no reason to rule this out. In an epidemic situation where each sample is taken from a different host, this known to be impossible, as there must have been at least one infection event since that TMRCA, and in the time from infection to sampling, a host will have gone through an incubation period and probably also a non-negligible period from manifestation of symptoms to sampling. If a single tree with these short terminal branch lengths is then used to estimate epidemiological parameters, estimates of times from infection to sampling are unlikely to be reliable.

Phylogenetic inference, too, would benefit from a more realistic population model for data from epidemics than the free mixing that is assumed in the standard coalescent-based methods. Much more sophisticated models, designed specifically with epidemics in mind, exist in the field of mathematical epidemiology. Of

particular interest are those [32, 51, 81] that treat each infected host or premises as an individual entity rather than the member of a compartment, as this aligns closely with phylogenetics, where each isolate must come from one particular host, and allows inference that uses detailed epidemiological data, which can be acquired at the same time that a pathogen sample is taken for sequencing.

I make two contributions here. Firstly, I provide a more rigorous mathematical definition of the correspondence identified by Didelot et al. [35] between an annotation of the internal nodes of a phylogeny with host data and a transmission tree. Secondly, I outline a full, flexible, Bayesian MCMC framework for "one-step", simultaneous estimation of transmission trees and phylogenies, which uses a model of the pathogen population that is consistent with host-to-host transmission during an epidemic, and can integrate relevant epidemiological data.

5.2 Transmission trees as partitions of the node sets of phylogenies

Suppose the set of all units infected in the epidemic (be they infected organisms or infected premises; I use the word "host" from here on for brevity) is $\mathbf{A} = \{a_1, \ldots, a_N\}$ and the set of isolates is $\mathbf{B} = \{B_1, \ldots, B_M\}$. Let $f : \mathbf{B} \to \mathbf{A}$ be a map taking an isolate to the host it was sampled from and assume f is surjective, in other words every host provides an isolate (which also implies $M \geq N$). Suppose also that there was no reinfection or superinfection, and that transmission is a complete bottleneck: only a single genetic variant enters the newly infected host upon transmission. Let $\mathcal{G}$ be a phylogeny describing the ancestral relationship of the members of $\mathbf{B}$, with branch lengths in units of time. It consists of two components:

- A rooted, binary tree G with a set $\mathbf{E}_G$ of M tips labelled, via a function g, with the elements of $\mathbf{B}$ and a set $\mathbf{I}_G$ of $M-1$ internal nodes. Let $\mathbf{N}_G = \mathbf{E}_G \cup \mathbf{I}_G$ be the complete set of nodes. Let $\mathbf{G_B}$ be the set of all such trees with tips allocated to isolates via g and isolates allocated to hosts via f. The map $d = f \circ g : \mathbf{E}_G \to \mathbf{A}$ labels each tip with a host. Call G the *topology* of the phylogeny.

- A length function $l : \mathbf{N}_G \to (0, \infty)$ that takes each non-root node of G to the difference in calendar time (in arbitrary units on a forwards timescale) between the time of the event represented by that node and the time of the event represented by its parent. The event represented by an element u of $\mathbf{E}_G$ (a tip) is the sampling of the isolate from the host corresponding to u's label; the event represented by an element v of $\mathbf{I}_G$ (an internal node) is the coalescence of the two lineages represented by v's two child branches; this occurs at the TMRCA of those lineages. In contrast to the convention in most phylogenetic methods, I do indeed define a nonzero $l(r)$ for the root node r of G. Its value is largely arbitrary, but it must be greater than any plausible value for the time between the event (generally a coalescence) represented by r and the infection event that seeded the entire outbreak.

The length function l allows us to also define a height function $h : \mathbf{N}_G \to [0, \infty)$ that takes each node to the difference in time between the event represented by that node and the time at which the last isolate was sampled. This map defines a backwards timeline for events on the whole tree whose zero point is the latter time.

A transmission tree on $\mathbf{A}$ is a rooted, directed tree with M nodes labelled with the elements of $\mathbf{A}$. If $\mathcal{N}$ is such a tree, it can be thought of as a map $\mathcal{N} : \mathbf{A} \to \mathbf{A} \cup \emptyset$ taking each host a_i to its infector or to $\emptyset$ if a_i is the first host, and I will use this notation henceforth. Let $\mathbf{\Pi_A}$ be the set of all transmission trees on $\mathbf{A}$. ($\mathbf{\Pi_A}$ has

cardinality N^{N-1} by Cayley's formula, as there are N^{N-2} such trees and N choices of root for each.) Take G to be a topology as above, describing the ancestry of $\mathbf{B}$ without meaningful branch lengths. We are interested in the set of transmission trees in $\mathbf{\Pi_A}$ that are consistent with the ancestry represented by G.

It is quite obvious (see figure 1.3b) that if each node in G is mapped to the host in which the corresponding pathogen lineage was present, then the transmission tree is known. We now wish to formally establish this link. This will allow us to stop dealing with the transmission tree as a separate entity, and instead treat it as a function applied to the internal nodes of G.

Definition 5.2.1. Let $\mathbf{\Omega}^{G,d}$ be the set of partitions of $\mathbf{N}_G$ such that:

- If $\mathcal{P} \in \mathbf{\Omega}^{G,d}$ and $p \in \mathcal{P}$ (such a p being a subset of $\mathbf{N}_G$), then the nodes in p and the edges between them form a connected subgraph of $\mathbf{N}_G$. For brevity say "p forms a connected subgraph of $\mathcal{G}$" henceforth.
- If $\mathcal{P} \in \mathbf{\Omega}^{G,d}$ and $p \in \mathcal{P}$, then $|\mathbf{E}_G \cap\ p| \geq 1$ and $|\{d(u) : u \in \mathbf{E}_G \cap\ p\}| = 1$. In other words, p contains at least one tip and all tips in p map to the same element of $\mathbf{A}$ under d. Each p corresponds to one and only one host.

Definition 5.2.2. $\mathbf{\Omega}^{G,d}$ may be empty if d is such that no partitions of this type exist. If it is not, say G is *compatible* with d.

For a fixed G, define a map $c : \mathbf{A} \to \mathbf{N}_G$ taking a host $a_i \in \mathbf{A}$ to the most recent common ancestor of all tips $u \in \mathbf{E}_G$ with $d(u) = a_i$. If $|d^{-1}(a_i)| = 1$, $c(a_i)$ is a tip.

Proposition 5.2.3. *G is not compatible with d if and only if there exist hosts $a_i, a_j \in \mathbf{A}$ such that $|d^{-1}(a_i)| \geq 2$ and $|d^{-1}(a_j)| \geq 2$ and either $c(a_i)$ is an ancestor of $c(a_j)$ but there exists $v \in d^{-1}(a_i)$ ($d^{-1}(a_i)$ being all tips corresponding to isolates taken from a_i) such that v is a descendant of $c(a_j)$, or $c(a_j)$ is an ancestor of $c(a_i)$ but there exists $v \in d^{-1}(a_j)$ such that v is a descendant of $c(a_i)$.*

Proof. For "if", if $c(a_i)$ is an ancestor of $c(a_j)$ and there exists such a v, then if $\mathcal{P}$ is a partition such that $p \in \mathcal{P}$ contains the whole of $d^{-1}(a_j)$ and forms a connected subgraph of G, then it must contain $c(a_j)$, and thus a partition element containing $c(a_i)$ and v cannot form a connected subgraph as a path from one to the other must intercept $c(a_j)$. Likewise if there is a $p \in \mathcal{P}$ containing $d^{-1}(a_i)$ that forms a connected subgraph, then it must contain $c(a_j)$ and hence any partition element containing $d^{-1}(a_j)$ cannot.

For "only if", if no such hosts exist, put a partial order on the $c(a_i)$ for all i such that $c(a_i) \preceq c(a_j)$ if $c(a_i)$ is a descendant of $c(a_j)$. Permute the $c(a_i)$ into a sequence $U = \{c(a_{o(1)}), \ldots, c(a_{o(N)})\}$ such that $o(i) \leq o(j)$ if $c(a_{o(i)}) \preceq c(a_{o(j)})$. Build a partition $\mathcal{P}$ by moving through U, assigning each $c(a_{o(i)})$ and each descendant of $c(a_{o(i)})$ that is not $c(a_{o(j)})$ or a descendant of $c(a_{o(j)})$ for $j < i$ to a new partition element. At the end of the process, perform a post-order traversal of the tree, assigning any remaining unassigned nodes encountered to a partition containing one of their children. It is clear that at the end, $\mathcal{P}$ has the required N elements. By construction, all nodes assigned to the same partition element form a connected subgraph of G, so it remains only to check the second half of definition 5.2.1. Suppose there exists an a_i such that there is a tip $v_i \in d^{-1}(a_i)$ which was not assigned to the same partition element as $c(a_i)$. This implies that there exists an a_j such that $c(a_j)$ is a descendant of $c(a_i)$ and v_i is a descendant of $c(a_j)$, which is the only way v_i would not have been assigned to $c(a_i)$'s element. As neither $c(a_i)$ nor $c(a_j)$ can be a tip, $|d^{-1}(a_i)| \geq 2$ and $|d^{-1}(a_j)| \geq 2$, and a_i and a_j are the kind of hosts we assumed did not exist. So all tips in each $d^{-1}(a_i)$ are assigned to the same element, the one containing $c(a_i)$, and this set of tips has size at least one since $c(a_i)$ is a common ancestor of at least one node. □

Corollary 5.2.4. *If $N = M$ or $N = M - 1$ then all topologies G are compatible with d.*

Proof. In this case all, or all but one, of the $c(a_i)$ are tips and thus have $|d^{-1}(a_i)| = 1$ □

From now on, assume that we are working with a G is compatible with d. For $\mathcal{P} \in \mathbf{\Omega}^{G,d}$, extend d to a map $d_{\mathcal{P}} : \mathbf{N}_G \to \mathbf{A}$ that takes each node of G to the host of the tips that are in the same element of $\mathcal{P}$ as itself. For each $a_i \in \mathbf{A}$, let $H_{\mathcal{P},i}$ be the subtree of G constructed by removing all nodes, and edges adjacent to them, that do not map to a_i under $d_{\mathcal{P}}$. Because $H_{\mathcal{P},i}$ is connected, it has a single root node. Define a second map $e_{\mathcal{P}} : \mathbf{A} \to \mathbf{N}_G$ taking each a_i to this root node. For brevity write $s_i = e_{\mathcal{P}}(a_i)$. All s_i have a parent s_iP in G, except for the root r of G (which must be the root of one such subtree).

We interpret a partition $\mathcal{P}$ in $\mathbf{\Omega}^{G,d}$ such that the lineages represented by all nodes in $\mathcal{P}$ were present in the single host that all tips in $\mathcal{P}$ were sampled from. Then $\mathcal{P}$ can be taken to a transmission tree by using $d_{\mathcal{P}}$ to annotate each node u of G with that host. We then know who infected whom; infection events occur along branches of G whose start and end nodes are in different elements of $\mathcal{P}$. The preimage of $a_i \in \mathbf{A}$ under $d_{\mathcal{P}}$ is the set of nodes of $H_{\mathcal{P},i}$. The rules by which partitions are defined correspond to the assumptions about the epidemic. The connectedness requirement implies no reinfection or superinfection (if a host could experience multiple infections then its corresponding partition element would be disconnected) and also that transmission is a complete bottleneck (or else the two child lineages of an internal node could both be transmitted to the same host at the same time, and again the partition element corresponding to that host would be disconnected). The requirement that all partition elements contain a tip is a result of the surjectivity of f (every host is sampled at least once). Proposition 5.2.3 shows that if G is not compatible with f, then the assumption of no reinfection or superinfection must be violated due to the placement of tips from the the same host in the phylogeny.

To formalise the correspondence, we construct a map $z : \mathbf{\Omega}^{G,d} \to \mathbf{\Pi}_{\mathbf{A}}$ such that if $\mathcal{P} \in \mathbf{\Omega}^{G,d}$ and $a_i \in \mathbf{A}$,

$$z(\mathcal{P})(a_i) = \begin{cases} d_{\mathcal{P}}(s_i P) & s_i \neq r \\ \emptyset & s_i = r \end{cases}$$

In other words, $z(\mathcal{P})$ returns the infector of a_i if the partition is $\mathcal{P}$.

Proposition 5.2.5. *For $\mathcal{P} \in \mathbf{\Omega}^{G,d}$, the directed graph T given by drawing an edge from $z(\mathcal{P})(a_i)$ to a_i for all $a_i \in \mathbf{A}$ is a directed tree, and if r is the root of G, the directionality is consistent with $d_{\mathcal{P}}(r)$ being the root of T.*

Proof. For the first part, we must show that the underlying undirected graph T' of T is connected and has no simple cycles. Suppose that it has a simple cycle passing through $n > 1$ distinct nodes $a_1, \ldots, a_n$. The construction of T will never give a node with indegree greater than 1 (as every host is infected once only), so this cycle must be directed in T; without loss of generality suppose the sequence $a_1, \ldots, a_n$ follows this directionality. Then $z(\mathcal{P})(a_k) = a_{k+1}$ for all $1 \leq k \leq n-1$ and $z(\mathcal{P})(a_n) = a_1$. If $i \geq 2$, $H_{\mathcal{P},i}$ is a subtree of G containing a root node s_i and the parent $s_{i-1}P$ of the root node of the subtree $H_{\mathcal{P},i-1}$; similarly $H_{\mathcal{P},1}$ contains $s_n P$. Since $H_{\mathcal{P},i}$ for each $i \geq 2$ contains a sequence of nodes, following the directedness of G induced by its root, running from s_i to $s_{i-1}P$, $H_{\mathcal{P},1}$ contains one running from s_1 to $s_n P$, and there is a directed link from each $s_i P$ to s_i in G, the concatenation of all of these forms a simple cycle in G, contradicting the fact it is a tree.

For connectedness, again suppose $a_i \in \mathbf{A}$ and let $a_j = d_{\mathcal{P}}(r)$; r is the root of both $H_{\mathcal{P},j}$ and G. It may be that $a_i = a_j$. If not, the path in G from s_i to s_j intersects $n \geq 2$ elements of $\mathcal{P}$ whose members map under $d_{\mathcal{P}}$ to the hosts $a_{o(1)}, \ldots, a_{o(n)} \in \mathbf{A}$, where o is some permutation of $\{1, \ldots, N\}$ with $o(1) = i$ and $o(n) = j$. In particular it must pass through the root nodes of all these

subtrees, $s_{o(1)}, \ldots, s_{o(n)}$, implying that $z(\mathcal{P})(a_{o(k)}) = a_{o(k+1)}$ for all $1 \leq k \leq n-1$. It follows that $(z(\mathcal{P}))^{n-1}(a_i) = a_j$; thus all elements of $\mathbf{A}$ are connected to a_j and furthermore to each other in T'. This also implies the existence of a directed path in T from a_j to any other a_i.

For the second part, $d_{\mathcal{P}}(r)$ has indegree 0 by construction, and we already have a directed path from $d_{\mathcal{P}}(r)$ to each $a \in \mathbf{A}$. As we have shown T is a tree, this is the only such path, hence the direction of all edges is away from $d_{\mathcal{P}}(r)$. □

Proposition 5.2.6. *z is injective.*

Proof. Suppose that there are two partitions $\mathcal{P}, \mathcal{P}'$ that have the same image under z, i.e. for all $a_i \in \mathbf{A}$, $z(\mathcal{P})(a_i) = z(\mathcal{P}')(a_i)$. If $\mathcal{P} \neq \mathcal{P}'$ then there exists some node u of G that has $a_i = d_{\mathcal{P}}(u) \neq a_j = d_{\mathcal{P}'}(u)$. It can be assumed that either u is the root of G or $d_{\mathcal{P}}(uP) = d_{\mathcal{P}'}(uP)$ for the parent uP of u (otherwise it is possible to move up G, towards the root, to find a new u for which this is true).

If u is the root of G, then it is the root of the subtrees $H_{\mathcal{P},i}$ and $H_{\mathcal{P}',j}$. This implies $z(\mathcal{P})(a_i) = \emptyset$ but $z(\mathcal{P}')(a_i) \neq \emptyset$ because $z(\mathcal{P}')(a_j) = \emptyset$ and only one element of $\mathbf{A}$ has image $\emptyset$ under $z(\mathcal{P}')$ since the root of G is unique. So uP exists.

Let $a_k = d_{\mathcal{P}}(uP) = d_{\mathcal{P}'}(uP)$. First suppose $k \neq i$ and $k \neq j$. Then $z(\mathcal{P})(a_i) = a_k$. We show that $z(\mathcal{P}')(a_i) = a_k$ is not possible. Let v be any tip of G with $d(v) = a_i$. Now v is a descendant of u because u is the root node of the subtree $H_{\mathcal{P},i}$, and $H_{\mathcal{P},i}$ includes v. $\mathcal{P}'$ gives rise to another subtree of G, $H_{\mathcal{P}',i}$, all of whose nodes map to a_i under $d_{\mathcal{P}'}$. This $H_{\mathcal{P}',i}$ has a root node s'_i which is *not* u because $d_{\mathcal{P}'}(u) = a_j$. It must, in fact, also be a descendant of u; if it were not, $H_{\mathcal{P}',i}$ would not be connected as it would include a node v that was a descendant of u and nodes that were not. The parent s'_iP cannot have $d_{\mathcal{P}'}(s'_iP) = a_k$ because either a) $s'_iP = u$ and $d_{\mathcal{P}'}(u) = a_j$ by construction or b) $s'_iP \neq u$ and if $d_{\mathcal{P}'}(s'_iP) = a_k$ were true, the set of nodes that map to a_k under $d_{\mathcal{P}'}$ would not be connected in G because they

would include uP which is u's parent and $s_i'P$ which is a descendant of u. Hence $z(\mathcal{P}')(a_i) \neq a_k$.

So without loss of generality suppose $k \neq i$ but $k = j$. Again $z(\mathcal{P})(a_i) = a_k$. Recall that $d^{-1}(a_k)$ is the set of tips of G that map to a_k under d and by extension $d_{\mathcal{P}}$ and $d_{\mathcal{P}'}$. No elements of $d^{-1}(a_k)$ are descendants of u. If any were, then $H_{\mathcal{P},k}$, the subgraph of G whose nodes are mapped to a_k by $d_{\mathcal{P}}$, would be disconnected by u, which maps to a_i. This implies that there is a node w of G, either a descendant of u or u itself, which maps to a_k under $d_{\mathcal{P}'}$ but neither of whose children wC_1 and wC_2 do. If $w = u$ then $d_{\mathcal{P}}(w) \neq a_k$ by construction. If $w \neq u$ and $d_{\mathcal{P}}(w) = a_k$, w would have an ancestor, u, which did not map to a_k under $d_{\mathcal{P}}$, and an earlier ancestor, uP, which did, breaking connectedness. This implies that $z(\mathcal{P}')(wC_1) = z(\mathcal{P}')(wC_2) = a_k$ but $z(\mathcal{P})(wC_1) = z(\mathcal{P})(wC_2) \neq a_k$.

□

For the next proposition, we need the following:

Lemma 5.2.7. *If $a_i, a_j \in \mathbf{A}$ and $\mathcal{N} \in \mathbf{\Pi_A}$ is a transmission tree in which a_i is an ancestor of a_j, then if $\mathcal{P} \in \mathbf{\Omega}^{G,d}$, $z(\mathcal{P}) = \mathcal{N}$, and u is a node of G with $d_{\mathcal{P}}(u) = a_j$, u has an ancestor v in G with $d_{\mathcal{P}}(v) = a_i$.*

Proof. Strong induction on the number n of intervening hosts between a_i and a_j in $\mathcal{N}$. If $n = 0$, this is true by definition of u, as the node s_j is an ancestor of u and its parent maps to a_i. If the lemma is true for all $n \leq m$ and the set of intervening hosts has size $m + 1$, let a_k be an arbitrary member of that set. The number of intervening hosts between a_k and a_j in $\mathcal{N}$ is less than $m + 1$, so u has an ancestor v in G with $d_{\mathcal{P}}(v) = a_k$. The number of intervening hosts between a_i and a_k in $\mathcal{N}$ is also less than $m + 1$, so v has an ancestor w in G with $d_{\mathcal{P}}(w) = a_i$. It follows that w is the ancestor of u needed. □

Proposition 5.2.8. *z is not surjective for $N > 2$.*

Proof. If $N > 2$ then $M > 2$. Let $a_i, a_j, a_k \in \mathbf{A}$ be any three hosts. In G, Let t_i, t_j, and t_k be any three tips with $d(t_i) = a_i$, $d(t_j) = a_j$ and $d(t_k) = a_k$. These tips have a most recent common ancestral node u and two of them, without loss of generality t_j and t_k, have a most recent common ancestral node v which is a descendant of u. We show that there is no element of $\mathbf{\Omega}^{G,d}$ which will map to any member of $\mathbf{\Pi_A}$ in which any of the following are true:

- a_j is an ancestor of a_i, which is an ancestor of a_k.
- a_j is an ancestor of a_k, which is an ancestor of a_i.
- a_k is an ancestor of a_i, which is an ancestor of a_j.
- a_k is an ancestor of a_j, which is an ancestor of a_i.

Let $\mathcal{P}$ be a partition such that $z(\mathcal{P})$ is a transmission tree in which a_j is an ancestor of both a_i and a_k (if no such transmission tree exists, then surjectivity is instantly disproven). Now $d_{\mathcal{P}}(u) = a_j$. To see this, note that since u is an ancestor of t_j, if it does not map to a_j under $d_{\mathcal{P}}$ then neither do any of its ancestors, because u would interrupt any path from t_j to such an ancestor and break connectedness. No descendants of the child of u which is not an ancestor of t_j and t_k map to a_j either, for the same reason, and this set includes t_i. All ancestors of t_i apart from u belong to one of those two categories. But lemma 5.2.7 is now contradicted because t_i has no ancestor which maps to a_j under $d_{\mathcal{P}}$ despite the fact that a_j is an ancestor of a_i in $z(\mathcal{P})$.

Now t_i has no ancestor in G that maps to a_k under $d_{\mathcal{P}}$, because the node u breaks connectedness between t_k and any position that such a node could be. The contrapositive of lemma 5.2.7 then says that a_k is not an ancestor of a_i in $z(\mathcal{P})$.

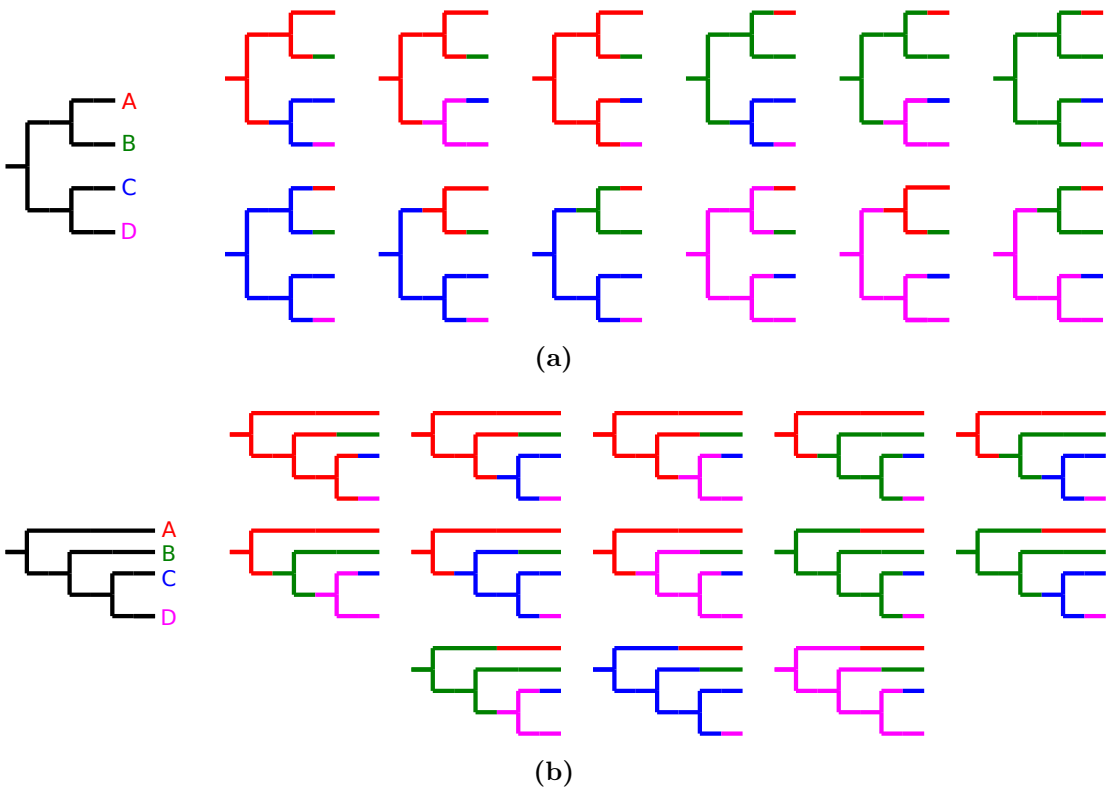


Figure 5.1: Illustration of the differing number of partitions of two phylogenies with the same tip count. a) the twelve valid partitions of the phylogeny ((A,B),(C,D)) for four hosts. b) the thirteen valid partitions of the phylogeny (A,(B,(C,D))) for four hosts

Similarly a_i is not an ancestor of a_k. An identical argument will show that if $z(\mathcal{P})$ is such that a_k is an ancestor of both a_i and a_k, a_i is not an ancestor of a_j nor vice versa.

□

Let the image of $\mathbf{\Omega}^{G,d}$ under z be $\mathbf{\Pi}_{\mathbf{A}}^{G,d} \subseteq \mathbf{\Pi}_{\mathbf{A}}$. The actual cardinality of $\mathbf{\Pi}_{\mathbf{A}}^{G,d}$ varies with the topology G, which can be clearly seen in the case $M = 4$ and $N = 4$ (figure 5.1).

Proposition 5.2.6 states that no two partitions of the internal nodes of G correspond to the same transmission history; the set of partitions and the set of transmission trees that are actually possible if G is the correct ancestry are equivalent. Proposition 5.2.8 shows, however, that not every possible transmission tree on $\mathbf{A}$ actually corresponds to a partition of the nodes of a fixed G, except in the trivial case where there are only two hosts. If we are interested in exploring the complete space of transmission trees using this construction, the phylogenetic topology must be varied as well.

Let the set $\mathbf{\Omega} = \{\mathbf{\Omega}^{G,d} : G \in \mathbf{G_B}\}$ consist of all partitions of all phylogenies with tips labelled with $\mathbf{B}$ (via a map g) and $\mathbf{A}$ (via $d = g \circ f$). The map z can be extended to a map $Z : \mathbf{\Omega} \to \mathbf{\Pi_A}$ in the obvious way.

Proposition 5.2.9. *Z is surjective. In other words, any transmission tree on $\mathbf{A}$ arises as a partition of some phylogenetic tree topology $G \in \mathbf{G_B}$.*

Proof. Let $\mathcal{N} \in \mathbf{\Pi_A}$. Use the following procedure to construct an element of $\mathbf{\Omega}$. For all i, suppose m_i is the number of isolates taken from $a_i \in \mathbf{A}$ (in other words, $m_i = |d^{-1}(a_i)|$) and a_i has n_i children in $\mathcal{N}$. Consider the phylogeny of the lineages infecting the host a_i. This has $m_i + n_i$ tips (one for each sample taken and one for each lineage that was transmitted to another host) and hence $m_i + n_i - 1$ internal nodes. However, the n_i tips corresponding to onwards infections do not represent nodes in the full phylogeny of the epidemic, so let $u^i_1, \ldots, u^i_{m_i}$ be nodes that are to represent sampling events (tips in the full phylogeny), and $v^i_1, \ldots, v^i_{m_i+n_i-1}$ be nodes to represent common ancestors.

Pick an arbitrary ordering of the children of each a_i in $\mathcal{N}$ and draw edges from each v^i_k to v^i_{k+1} and from v^i_k to v^j_1 where j and k are such that a_j is the kth child of a_i in the ordering. For each i, the nodes $v^i_1, \ldots, v^i_{n_i}$ now have two children each; $v^i_{n_i+1}, \ldots, v^i_{n_i+m_i-1}$ still have none. There are $m_i - 1$ of those, so they and $u^i_1, \ldots, u^i_{m_i}$ can be connected into an arbitrary binary tree with the former as

internal nodes, the latter as tips, and $v^i_{n_i+1}$ (which is already connected to $v^i_{n_i}$ by an edge) as the root. When this has been performed for all i, call the full graph G. If $l \in \{1, \ldots, N\}$ is such that a_l is the root of $\mathcal{N}$, let the root of G be $v_{l,1}$.

It is clear that G is a rooted binary tree. Its tip set $\mathbf{E}_G$ consists of $u^i_1, \ldots, u^i_{m_i}$ for each i. Let g be any bijective map from $\mathbf{E}_G$ to $\mathbf{B}$ such that $f \circ g(u^i_j) = a_i$ for all i and j. For each i, the set of nodes $\{u^i_1, \ldots, u^i_{m_i}\} \cup \{v^i_1, \ldots, v^i_{m_i+n_i-1}\}$ forms, by construction, a connected subtree of G and contains a nonempty set of tips whose image under $d = f \circ g$ is of size one; hence this partition of the nodes of G is an element $\mathcal{P}$ of $\mathbf{\Omega}^{G,d}$. It is easily checked that $z(\mathcal{P}) = \mathcal{N}$. □

As an aside, the arbitrary choices made in this construction imply that Z is clearly not injective in general, or in other words, two partitions of different phylogenies can correspond to the same transmission tree. (In fact, some elements of $\mathbf{\Omega}$ cannot be produced by the construction of proposition 5.2.9 at all, for example, the bottom right example in figure 1.3b if $N = M = 3$.) The upshot of proposition 5.2.9 is that a MCMC procedure that fully explores the space of these partitioned phylogenies is also fully exploring the space of transmission trees amongst the elements of $\mathbf{A}$. I outline such a procedure in section 5.3.

So far, I have only dealt with the topology G of the phylogeny. If this construction is to be useful for epidemic reconstruction, branch lengths must also be considered. Let $\mathcal{P}$ be a partition of G, and suppose G is the topology of a phylogeny $\mathcal{G}$ with length function l and height function h. Suppose $a_i \in \mathbf{A}$ and that $z(\mathcal{P})(a_i) \neq \emptyset$. Let $u = e_{\mathcal{P}}(a_i)$ (the root of the subtree mapped to a_i by $d_{\mathcal{P}}$), and let uP be the parent of u, which if it exists must be in a different partition element. An infection event occurs on the branch between uP and u, which means, assuming that internal nodes of G and transmissions do not occur at exactly the same time, that it occurs at a height in the open interval $(h(u), h(uP))$. It is more convenient to use a forwards timescale, so let $C : \mathbb{R} \to \mathbb{R}$ be a function converting

between tree height and such a timescale (in the same units, so branch lengths are maintained). Let t_i^{inf} be this time, on the forwards scale, of this infection event. Let $q_i \in (0,1)$ be such that $t_i^{\text{inf}} = C(h(uP)) + q_i l(u)$. If uP does not actually exist, i.e. a_i is the first host in the epidemic, then t_i^{inf} is between $C(h(r) + l(r))$ and $C(h(r))$ (remembering that we gave the root r of G a finite branch length) we can similarly define q_i such that $t_i^{\text{inf}} = C(h(r) + l(r)) + q_i l(r)$.

The combination of a phylogeny $\mathcal{G}$, map f from tip set to host set, partition $\mathcal{P}$ and a set of q_is for all elements $a_i \in \mathbf{A}$ then entirely determines the transmission history of the epidemic, describing which host infected which others and when. No assumptions are made at this, conceptual, level about when hosts cease to be infectious; a host can continue to infect others at any time following the time at which a sample was acquired from it. If, as will often be the case, this is an unreasonable assumption, the likelihood of such partitions can be evaluated to zero in the calculation of the posterior probability.

5.3 MCMC procedure

I showed in section 5.2 that, if the sequence data is such that at least one sample is taken from each host, every transmission tree arises as at least one member of the set of partitioned phylogenies. Thus a Bayesian MCMC procedure to estimate time-resolved phylogenies can be extended to one that simultaneously samples from the probability distribution of reconstructed epidemics if each sampled tree $\mathcal{G}$ is augmented with a partition of its internal nodes as well as a set of values $\{q_i\}$ determining the exact times of infection. (An alternative approach, which I do not employ here but may be essential in extending the procedure to accommodate unsampled hosts, would be to insert an internal, binary node to represent each transmission event.) Because of the special requirements of this

type of augmentation, the standard MCMC moves on a phylogenetic tree topology used by a package such as BEAST are unsuitable as they will generally not make modifications that respect the rules of the node partitions. Instead, a specialised set moves have been devised to alter the phylogeny and partition in such a way that the transmission tree structure is maintained, which I now describe.

Note that these moves do not simultaneously change the value of any of the q_is, as new values of these are proposed and evaluated separately. Nevertheless, changes to either tree may involve resampling the times of infection of some hosts. If $a_i \in \mathbf{A}$, changing partition from $\mathcal{P}$ to $\mathcal{P}'$ may mean that $e_{\mathcal{P}}(a_i)$ and $e_{\mathcal{P}'}(a_i)$, the roots of the partition elements corresponding to a_i, are different nodes with different heights, and so while q_i will not change, the time of infection t_i^{inf} of a_i will. Even a move that has no effect on the partition or phylogenetic tree topology, such as a change to branch lengths, may also alter the height of $e_{\mathcal{P}}(a_i)$ and/or its parent, which will also modify t_i^{inf} while q_i remains fixed.

For the following definitions, recall that, for a host a_i and a partition $\mathcal{P}$ of a phylogenetic tree topology G, $H_{\mathcal{P},i}$ is the subtree of G whose nodes are mapped to a_i under $d_{\mathcal{P}}$, $e_{\mathcal{P}}(a_i)$ is the root of this subtree, and $c(a_i)$ is the MRCA of all tips corresponding to isolates sampled from a_i (which may be $e_{\mathcal{P}}(a_i)$ and otherwise is descended from it).

Definition 5.3.1. For a partition $\mathcal{P}$ of a phylogeny $\mathcal{G}$ with topology G determining a transmission tree on a set $\mathbf{A}$ of hosts, if u is a phylogenetic tree node with $d_{\mathcal{P}}(u) = a_i \in \mathbf{A}$ say u is *ancestral under* $\mathcal{P}$ if it is an ancestor of a node of the subtree $H_{\mathcal{P},i}$ which is also a tip of G. To put it another way, there is a tip v of G that is mapped to a_i by $d_{\mathcal{P}}$ such that it is possible to draw a simple path from v to the root of G that passes through u.

Definition 5.3.2. For a partition $\mathcal{P}$ of a phylogeny $\mathcal{G}$ with topology G determining

a transmission tree on a set $\mathbf{A}$ of hosts, the *infection branch* for $a_i \in \mathbf{A}$ is the branch of G ending in $e_{\mathcal{P}}(a_i)$.

Definition 5.3.3. For a phylogeny $\mathcal{G}$ whose topology G is compatible with a map d taking each tip to the host of the corresponding isolate, say $a_i \in \mathbf{A}$ is *root-blocked* by $a_j \in \mathbf{A}$ if $c(a_j)$ is an ancestor of $c(a_i)$.

These definitions are illustrated in figure 5.2a. It should be noted that:

- For any valid partition $\mathcal{P}$, $d_{\mathcal{P}} \circ c(a_i)$ is a_i itself.
- If $N = M$, i.e. there is only one isolate per host, then $c(a_i)$ is the unique tip whose isolate was sampled from a_i for all i.
- As a result, if $N = M$ then no hosts are root-blocked by any others as all $c(a_i)$s are tips.
- If a_i is root-blocked by any a_j then the root r of G cannot be in the partition containing $d^{-1}(a_i)$, because $c(a_j)$ must lie on the path from $c(a_i)$ to r and for any $\mathcal{P}$, $d_{\mathcal{P}} \circ c(a_i) \neq d_{\mathcal{P}} \circ c(a_j)$ if $i \neq j$, so connectedness would be violated.

Suppose $\mathcal{G}$ is a phylogeny with tree topology G and $\mathcal{P} \in \mathbf{\Omega}^{G,d}$ a partition of its nodes according to definition 5.2.1. In what follows, trees are oriented so the "up" direction is towards the root and the "down" towards the tips.

5.3.1 Infection branch operator

Randomly select a host a_i that is not the first host in the outbreak (i.e. $e_{\mathcal{P}}(a_i)$ is not the root of G). Let $u = e_{\mathcal{P}}(a_i)$ and uP be the parent of u (which must exist as the root was avoided). The operator performs both "downward" and "upward"

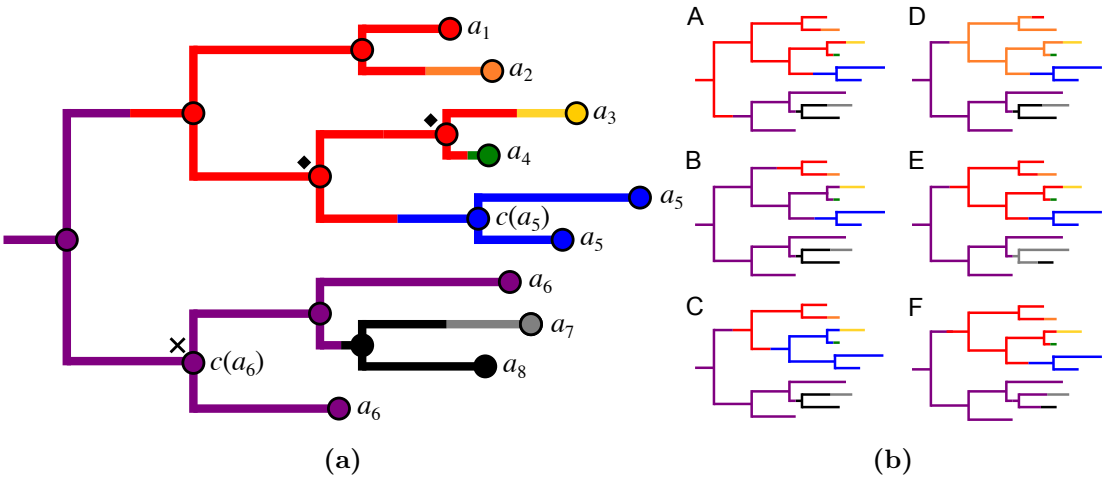


Figure 5.2: Illustration of a partitioned phylogeny and the behaviour of the infection branch operator. a) A partitioned phylogeny; nodes (circles) are coloured by the partition element they belong to. Tips are labelled by the hosts that the isolates corresponding to them were taken from. Where more than one isolate is taken from a host a_i, $c(a_i)$ is labelled; in all other cases $c(a_i)$ is the single tip corresponding to an isolate taken from that host. Black diamonds designate nodes that are not ancestral under the partition. The hosts a_7 and a_8 are root-blocked by a_6 due to the position of $c(a_6)$ (black cross). b) Some effects of performing the infection branch move on this partition. (A) Upward move on a_1; note the change of the partition element the root node belongs to. (B) Downward move on a_1. (C) Upward move on a_5. (D) Upward move on a_2. (E): Upward move on a_7. (F): Downward move on a_8.

moves, but if $u = c(a_i)$ (which is true if u is a tip) then the move must be upwards and if both a) $d_{\mathcal{P}}(u)$ is root-blocked by $d_{\mathcal{P}}(uP)$ and b) uP is ancestral under $\mathcal{P}$ then the move must be downwards; if both are true the move fails. In other cases, upwards or downwards are each selected with probability 0.5. It must be that u and uP are in different elements of $\mathcal{P}$, and this implies that u is ancestral under $\mathcal{P}$ because the path from any node v that is not a descendant of u to u must pass through uP and if $d_{\mathcal{P}}(v) = a_i$ this would violate the connectedness requirement. Suppose $d_{\mathcal{P}}(uP) = a_j$.

Downward move Propose a new partition $\mathcal{P}'$ that has $d_{\mathcal{P}'}(u) = a_j$, moving the infection branch of a_i down the tree. Consider the two children uC_1 and uC_2 of u (as this is the downward move, u is not a tip). At least one of these is mapped to the same element of $\mathbf{A}$ as u by $d_{\mathcal{P}}$ because u must be in the same element of $\mathcal{P}$ as $c \circ d_{\mathcal{P}}(u)$ and the path from u to this node in the subtree will intersect one of its children. If this is true of only one child then without loss of generality say it is uC_1. In this case simply make $\mathcal{P}'$ by setting $d_{\mathcal{P}'}(i) = a_j$ and leave the rest of the partition unchanged; this is clearly still a valid partition because all subtrees remain connected. So suppose also $d_{\mathcal{P}}(uC_2) = d_{\mathcal{P}}(u)$. One and only one of uC_1 and uC_2 is ancestral under $\mathcal{P}$ (they would only both be if $u = c(a_i)$ which was prohibited and if neither is, the subtree $H_{\mathcal{P},i}$ is either not connected or contains no tip) so, again without loss of generality, say it is uC_1. If we again set $d_{\mathcal{P}'}(u) = a_j$, the removal of u from $H_{\mathcal{P},i}$ splits the nodes of the latter into two sets, V_1 containing uC_1 and $c \circ d_{\mathcal{P}}(u)$, and V_2 containing uC_2. The nodes of both sets and the edges between them form connected subtrees of T, but their union is not connected. Complete the construction of $\mathcal{P}'$ by setting $d_{\mathcal{P}'}(v) = a_j$ for all $v \in V_2$. $H_{\mathcal{P}',i}$ and $H_{\mathcal{P}',j}$ are then connected.

The effect on the transmission tree is that all $a_k \in \mathbf{A}$ that have $z(\mathcal{P})(a_k) = a_i$ and $c(a_k)$ a descendant of or equal to uC_2 have $z(\mathcal{P}')(a_k) = a_j$ instead.

Upward move Propose a new partition $\mathcal{P}'$ that has $d_{\mathcal{P}'}(uP) = a_i$, moving the infection branch of a_i up the tree. We need to consider the grandparent uG of u if it exists, and the sibling uS of u (the other child of uP). At least one of uG and uS must be in the same element of $\mathcal{P}$ as uP (or else uP is not in a partition element containing a tip). If uG does not exist then this must be uS.

If $d_{\mathcal{P}}(uS) = a_j$ and either $d_{\mathcal{P}}(uG) \neq a_j$ or uG does not exist, then setting $d_{\mathcal{P}'}(uP) = a_i$ is all that is required to make $\mathcal{P}'$ a valid partition. The two or three nodes joined to uP by edges were all in different elements of $\mathcal{P}$ and remain so; uP was in the element of $\mathcal{P}$ containing one of its children and is moved to the one containing the other child in $\mathcal{P}'$. Similarly, if $d_{\mathcal{P}}(uG) = a_j$ and $d_{\mathcal{P}}(uS) \neq d_{\mathcal{P}}(uP)$, then all that is necessary is to set $d_{\mathcal{P}'}(uP) = a_i$; the situation is the same except that the uP has moved from the element of $\mathcal{P}$ that contains its parent to one containing one of its children.

If uG exists and $d_{\mathcal{P}}(uS) = d_{\mathcal{P}}(uG) = a_j$, then the removal of uP from the subtree $H_{\mathcal{P},j}$ splits the latter into two subtrees whose union is again not a connected subtree of G. Let the node sets of these two subtrees be V_1 and V_2, with V_1 containing uG and V_2 containing uS. V_1 and V_2 cannot both contain tips, because if they did, uP would be ancestral under $\mathcal{P}$ and $d_{\mathcal{P}}(u)$ would be root-blocked by $d_{\mathcal{P}}(uP)$ as $c(a_i)$ must be a descendant of u and $c(a_j)$ must be an ancestor of uP. If uP is ancestral under $\mathcal{P}$ then V_2 contains tips, and if it is not then V_1 does. Complete $\mathcal{P}'$ by setting $d_{\mathcal{P}'}(v) = a_i$ for all v in the set that contains no tips. $H_{\mathcal{P}',i}$ and $H_{\mathcal{P}',j}$ are now connected. Note that V_1 may contain the root node and if it does not contain $c(a_j)$ then the root's image under $d_{\mathcal{P}}$ is different from that under $d_{\mathcal{P}'}$, which is how this move may change the first host in the outbreak even though the root host is never chosen. This can be seen in example (A) of figure 5.2b.

If uP is not ancestral under $\mathcal{P}$, then the effect on the transmission tree is that all $a_k \in \mathbf{A}$ that have $z(\mathcal{P})(a_k) = a_j$ and $c(a_k)$ a descendant of or equal to uS have

$z(\mathcal{P}')(a_k) = a_i$ instead. If uP is ancestral under $\mathcal{P}$ then, in $z(\mathcal{P}')$, a_i is the infector of a_j instead of vice versa, and all $a_k \in \mathbf{A}$ that have $z(\mathcal{P})(a_k) = a_j$ and $c(a_k)$ *not* a descendant of uS have $z(\mathcal{P}')(a_k) = a_i$ instead.

Hastings ratio In every case the Hastings ratio is either 1, 2 or 1/2 depending on the exact nature of the nodes involved. We observe that:

- The downward move on u is reversed by the upward move on the child uC_1 of u that is ancestral under $\mathcal{P}$. The Hastings ratio is 1 multiplied by 2 if $uC_1 = c \circ d_{\mathcal{P}'}(uC_1)$ and then by 1/2 if uP is ancestral under $\mathcal{P}$ and $d_{\mathcal{P}}(u)$ is root-blocked by $d_{\mathcal{P}}(uP)$.
- If uP is not ancestral under $\mathcal{P}$, then the upward move on u is reversed by the downward move on uP. The Hastings ratio is 1 multiplied by 1/2 if $u = c \circ d_{\mathcal{P}}(u)$ and then by 2 if uG is ancestral under $\mathcal{P}$ and $d_{\mathcal{P}'}(uP)$ is root-blocked by $d_{\mathcal{P}'}(uG)$.
- If uP is ancestral under $\mathcal{P}$, and the upward move on u is possible, then it is reversed by the upward move on its sibling uS. The Hastings ratio is 1 multiplied by 1/2 if $u = c \circ d_{\mathcal{P}}(u)$ and then by 2 if $uS = c \circ d_{\mathcal{P}'}(uS)$.

The various effects of applying this move to a partitioned phylogeny are illustrated in figure 5.2b. It serves the same purpose as that proposed by Didelot et al. [35], but takes a rather different approach. The main difference it that this is a change to the tree partition, which may indirectly change an infection date, rather than to the infection dates themselves. Direct changes to infection dates in my framework are constrained to be those that cannot change the transmission tree, as they modify just the q_is. Other differences are that my version makes only moves that respect the partition rules (and hence the proposal never violates them), makes no assumption that cases cease to be infectious at any point (which is left as a

job for likelihood calculations) and also allows for multiple tips to correspond to isolates from the same host.

5.3.2 Phylogenetic tree operators

I have adapted the three standard tree moves used in BEAST (exchange, subtree slide, and Wilson-Balding [38, 67, 166]) such that they respect the transmission tree structure induced by partitioning the internal nodes. I give two versions of each:

- A "type A" operator which does not alter the transmission tree at all; all parental relationships remain the same.
- A "type B" operator which performs phylogenetic tree modifications which simultaneously rearrange the transmission tree by assigning new parents to one or two hosts.

For convenience, assume that the nodes of the phylogeny $\mathcal{G}$ are uniquely labelled. When $\mathcal{G}$ is modified to a proposed phylogeny $\mathcal{G}'$, it retains the same node set but has a different edge set. It is then meaningful for a single partition $\mathcal{P}$ to apply to the nodes of both $\mathcal{G}$ and $\mathcal{G}'$.

The previous work of Ypma et al. [173] on this topic treated each within-host phylogeny as a completely separate entity, to be updated independently. The transmission tree was modified by a single move similar (although not identical) to the Wilson-Balding type B proposal described here. Instead, I work with and modify the phylogeny for the whole epidemic, providing a suite of moves comparable to those already widely used in major phylogenetics packages. Irreducibility was also not proved in that paper and it is not immediately obvious that it holds;

the partition structure outlined here and also used by Didelot et al. [35] is mathematically much less complicated and irreducibility is (in both cases) quite straightforward to show. Another key difference is that I have provided a rigorous methodology based solely on the partitioned tree space; none of these moves checks any epidemiological information as part of the proposals, unlike both earlier examples. This work is instead expected to be done as part of likelihood calculations. This increases flexibility; the likelihood of a partitioned tree can be calculated according to a wide range of epidemiological models, but these tree moves will never have to be rewritten in order to accommodate such changes. There may also be uses for this partitioned space other than epidemic reconstruction; as previously noted [173], there are similarities between transmission trees and species trees in their relationship to phylogenies, with the main difference being that while it is important when considering transmission to consider who infected whom, a species is generally modelled as simply splitting into two. Some methods for simultaneous reconstruction of species trees and phylogenies simply reject proposals in which the two trees are not compatible [62]; here I have developed a suite of moves that do not ever propose such states when the "higher level" tree is a transmission tree. Adaptation may be possible to the realm of species trees as well.

The type A moves are illustrated in figure 5.3 and the type B moves in figure 5.4

Type A operators

Type A exchange Select a random node u that is not the root r of the phylogeny $\mathcal{G}$, and then randomly select a second node v, also not r and not the sibling uS of u, such that the parents uP and vP of u and v are in the same element of $\mathcal{P}$, $h(uP) > h(v)$, and $h(vP) > h(u)$ (recall that the height function is in backwards time from the last sample date). The last condition rules out the possibility

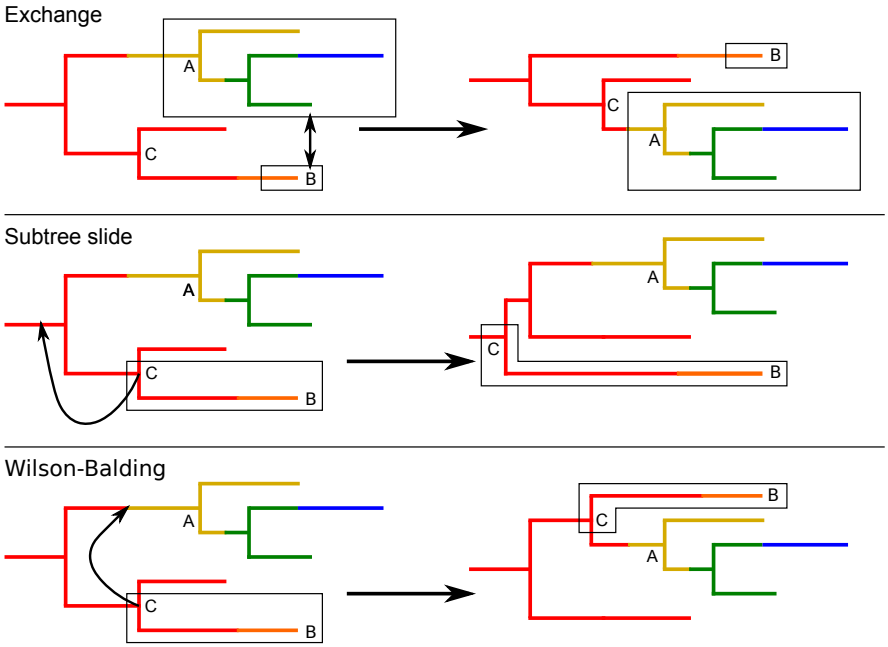


Figure 5.3: Depiction of the type A phylogeny operators. The exchange move exchanges the nodes marked A and B; the subtree slide and Wilson-Balding moves change the position of the node B and its parent C.

that u is the ancestor of v or vice versa. If there is no such v then the operator fails. Otherwise, u and v exchange parents to obtain a proposed phylogenetic tree $\mathcal{G}'$ with the same partition of nodes $\mathcal{P}$. To see that $\mathcal{P}$ is still valid in terms of connectedness, note that the only nodes which are adjacent to different nodes before and after the move are u, uP, v, and vP. If anything has been disconnected it must have been along those branches. But if $d_{\mathcal{P}}(u) \neq d_{\mathcal{P}}(uP)$ then there was already a partition change along the branch from u to uP without the rules being violated, so if there is one on the branch that is now from u to vP then the rules still hold; no path from u to any other member of its partition element has been modified. If, on the other hand, $d_{\mathcal{P}}(u) = d_{\mathcal{P}}(uP)$ then changing u's parent to vP means that it is still adjacent to a node with the same image under $d_{\mathcal{P}}$ as itself, and nothing has occurred to prevent there being a path between any two nodes in u's partition element. In both cases the same goes for v. The transmission tree structure is unchanged: if $d_{\mathcal{P}}(u) \neq d_{\mathcal{P}}(uP)$ then $d_{\mathcal{P}}(u)$ is infected by $d_{\mathcal{P}}(uP)$ before

the move and by $d_{\mathcal{P}}(vP) = d_{\mathcal{P}}(uP)$ afterwards, whereas if $d_{\mathcal{P}}(u) = d_{\mathcal{P}}(uP)$ then $d_{\mathcal{P}}(u)$'s infection branch was not affected at all. Again, the same goes for v.

For the Hastings ratio, note that the partitioned tree obtained by selecting u and then v is exactly the same as that obtained by selecting v and then u. If u is selected first, let $n^{\mathrm{EA}}_{\mathcal{G},\mathcal{P}}(u)$ be the number of eligible nodes to be selected second (this is explicitly calculated every time the operator acts). The node u is selected first with probability $\frac{1}{2M-2}$ and then v is selected with probability $\frac{1}{n^{\mathrm{EA}}_{\mathcal{G},\mathcal{P}}(u)}$. The outcome is the same if v is selected first with probability $\frac{1}{2M-2}$ and then u with probability $\frac{1}{n^{\mathrm{EA}}_{\mathcal{G},\mathcal{P}}(v)}$. The denominator of the Hastings ratio is thus $\frac{1}{2M-2}\left(\frac{1}{n^{\mathrm{EA}}_{\mathcal{G},\mathcal{P}}(u)} + \frac{1}{n^{\mathrm{EA}}_{\mathcal{G},\mathcal{P}}(v)}\right)$. The move is reversed by selecting the same two nodes again (in either order), hence $n^{\mathrm{EA}}_{\mathcal{G}',\mathcal{P}}(u)$ and $n^{\mathrm{EA}}_{\mathcal{G}',\mathcal{P}}(v)$ are calculated and the ratio's numerator is $\frac{1}{2M-2}\left(\frac{1}{n^{\mathrm{EA}}_{\mathcal{G}',\mathcal{P}}(u)} + \frac{1}{n^{\mathrm{EA}}_{\mathcal{G}',\mathcal{P}}(v)}\right)$.

Type A subtree slide Select a random node u under the conditions that $u \neq r$ and at least one of u's grandparent uG or sibling uS is in the same element of $\mathcal{P}$ as its parent uP. Draw a distance $\Delta \in \mathbb{R}$ from some probability distribution that is symmetric about 0. The move aims to change the height of uP to $h(uP) + \Delta$. If $\Delta > 0$, find the node v amongst uS and its ancestors which has the minimum height while fulfilling $h(v) < h(uP) + \Delta$; this may be the root node or uS itself. If v is not in the same element of $\mathcal{P}$ as uP then the move fails. If $v = uS$ then simply change the height of uP to $h(uP) + \Delta$ and the topology is unchanged. Otherwise, modify the tree such that uP has height $h(uP) + \Delta$, parent vP (or no parent if $v = r$ in which case uP is now the root node) and child v, and uS has parent uG. Again, do not change $\mathcal{P}$. Connectedness rules are still obeyed because, in the new tree $\mathcal{G}'$, uP is adjacent to v, which is in the same element of $\mathcal{P}$ as itself. The transmission tree structure is unchanged as:

- The move does not change the partition, so no infection branch has changed if the corresponding phylogenetic tree branch was not modified by the move, except possibly by changing its length. This applies to the branch between u and uP as well as all branches adjacent to nodes other than u, uP, uG, uS, v, and vP.

- If uS and uP are in different elements of $\mathcal{P}$ then uP and uG are in the same one, so the infector of $d_{\mathcal{P}}(uS)$ remains the same.

- If uG and uP are in different elements of $\mathcal{P}$ then the move would have failed if $h(uP) + \Delta > h(uG)$ so the phylogenetic tree topology is unchanged.

- If v and vP are in different elements of $\mathcal{P}$ then uP, instead of v, is now the top end of $d_{\mathcal{P}}(uP)$'s infection branch, but $d_{\mathcal{P}}(uP) = d_{\mathcal{P}}(v)$ and its infector is still $d_{\mathcal{P}}(vP)$.

If $\Delta < 0$, then if $h(uP) + \Delta < h(u)$ the move fails. Otherwise, the move selects a node v at random with equal probability from the set W which consists of nodes w that:

1. Are descendants of uP but not descendants of u.

2. Have $h(w) < h(uP) + \Delta$ but $h(wP) > h(uP) + \Delta$; i.e. height $h(uP) + \Delta$ occurs along the branch which it terminates.

3. Have a parent in the same partition element as uP.

If W is empty the move fails. In the case that W consists only of uS then simply set $h(uP) = h(uP) + \Delta$ and the topology is unchanged. Otherwise, modify the tree such that uP has height $h(uP) + \Delta$, parent vP and child v, and uS has parent uG. Connectedness rules are still obeyed because there is an edge from uP to a

node (vP) in the same element of the partition. The transmission tree structure is unchanged as:

- Again, the move does not change the partition, so any infection branches have not changed if the particular phylogenetic tree branch was not modified by the move, except by a change of length.
- If uS and uP are in different elements of $\mathcal{P}$ then the move would have failed if $h(uP) + \Delta < h(uS)$ so the topology is unchanged.
- If uG and uP are in different elements of $\mathcal{P}$ then uP and uS are in the same one, so the infector of $d_{\mathcal{P}}(uP)$ remains the same; uS is now the end of its infection branch.
- If v and vP are in different elements of $\mathcal{P}$ then the infector of $d_{\mathcal{P}}(v)$ is still $d_{\mathcal{P}}(vP) = d_{\mathcal{P}}(uP)$.

Suppose there are $n^{\mathrm{SA}}_{\mathcal{G},\mathcal{P}}$ nodes eligible for this move before it occurs and $n^{\mathrm{SA}}_{\mathcal{G}',\mathcal{P}}$ afterwards. If the topology did not change then the Hastings ratio is $\frac{n^{\mathrm{SA}}_{\mathcal{G},\mathcal{P}}}{n^{\mathrm{SA}}_{\mathcal{G}',\mathcal{P}}}$. Otherwise, it is $\frac{|W|n^{\mathrm{SA}}_{\mathcal{G},\mathcal{P}}}{n^{\mathrm{SA}}_{\mathcal{G}',\mathcal{P}}}$ if $\Delta < 0$ and $\frac{n^{\mathrm{SA}}_{\mathcal{G},\mathcal{P}}}{|W'|n^{\mathrm{SA}}_{\mathcal{G}',\mathcal{P}}}$ if $\Delta > 0$, where W' is the set of nodes w that:

1. Are descendants of vP (in $\mathcal{G}$) but not descendants of u.
2. Have $h(w) < h(uP)$ but $h(wP) > h(uP)$.
3. Have $d_{\mathcal{P}}(wP) = d_{\mathcal{P}}(v)$.

Type A Wilson-Balding move Pick a node u under the same conditions as for the type A subtree slide: $u \neq r$ and at least one of u's grandparent uG and sibling uS is in the same element of $\mathcal{P}$ as its parent uP. Pick a second node v

at random from amongst all nodes that are in the same element of $\mathcal{P}$ as uP, or whose parents are, and such that $h(vP) > h(u)$. The move fails if $uP = vP$, or $v = uP$. The node uP is pruned and reattached as a child of vP and the parent of v as with the standard Wilson-Balding move [38, 166]. As before, do not change $\mathcal{P}$. Connectedness rules are obeyed because there is an edge from uP to a node (either v or vP) in the same element of $\mathcal{P}$ as itself. The transmission tree structure is unchanged because if there was an infection event between uG and uC (and there was at most one by construction) then there still is and it involves the same hosts, and likewise if there was one between vP and v then there still is and it involves the same hosts. If there was no infection event in either case then the removal or insertion of uP does not add one.

Notice that if u is subsequently selected for this move again, then the set of candidates for the second node is the same except that it excludes v but includes uS; in particular it has the same cardinality, as it did for the standard Wilson-Balding move. So only the choice of first node affects the Hastings ratio. It follows that this is the ratio from the standard Wilson-Balding move multiplied by $\frac{n^{\mathrm{WA}}_{\mathcal{G},\mathcal{P}}}{n^{\mathrm{WA}}_{\mathcal{G}',\mathcal{P}}}$, where $n^{\mathrm{WA}}_{\mathcal{G},\mathcal{P}}$ is the number of nodes eligible for this move before it occurs and $n^{\mathrm{WA}}_{\mathcal{G}',\mathcal{P}}$ is the number afterwards.

Type B operators

Type B exchange Select a random node u, not r, whose parent uP is in a different element of $\mathcal{P}$ to itself. Pick a second node v, also not r and not uS, whose parent vP is also in a different element of $\mathcal{P}$ to itself (but this time the elements containing uP and vP do not have to be the same), such that $h(uP) > h(v)$, and $h(vP) > h(u)$, which as before prevents any ancestral relationship between u and v. If there is no such v then the operator fails. Otherwise, u and v exchange parents as with the type A operator to produce a proposal phylogeny $\mathcal{G}'$. $\mathcal{P}$ again

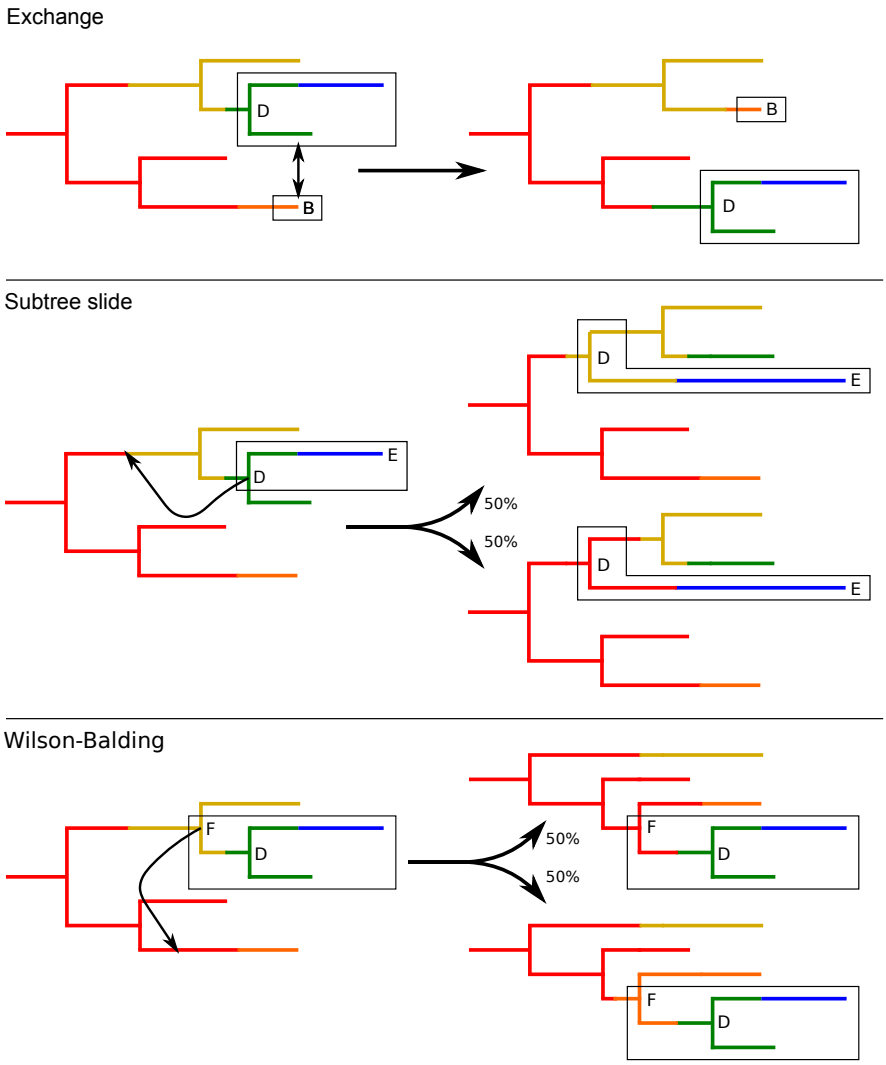


Figure 5.4: Depiction of the type B phylogeny operators. The exchange move exchanges the nodes marked B and D; the subtree slide move the node E and its parent D, and the Wilson-Balding the node D and its parent F. After the latter two moves the transplanted parent node is randomly assigned to one of two new partition elements with equal probability.

does not change. That it preserves connectedness of subtrees is clear; it does not change where the boundaries between partition elements occur at all. The effect on the transmission tree is that $d_{\mathcal{P}}(u)$ and $d_{\mathcal{P}}(v)$ exchange parents (if their parents are different).

The Hastings ratio is calculated in effectively the same way as for the type A version, noting that the number of choices for u is just $N-1$. If $n^{\mathrm{EB}}_{\mathcal{G},\mathcal{P}}(u)$ is the number of eligible choices for a second node if u is chosen first, then the ratio is $\left(\frac{1}{n^{\mathrm{EB}}_{\mathcal{G}',\mathcal{P}}(u)}+\frac{1}{n^{\mathrm{EB}}_{\mathcal{G}',\mathcal{P}}(v)}\right)/\left(\frac{1}{n^{\mathrm{EB}}_{\mathcal{G},\mathcal{P}}(u)}+\frac{1}{n^{\mathrm{EB}}_{\mathcal{G},\mathcal{P}}(v)}\right)$.

Type B subtree slide This time, u is a random node whose parent exists and is in a different element of $\mathcal{P}$ to itself. This implies that uP is in the same element as either uS or uG (if the latter exists) because otherwise uP would not be in a partition element containing a tip. The operator performs the standard subtree slide move [67] on u, by drawing a $\Delta \in \mathbb{R}$ from a probability distribution symmetric around 0, finding a node v such that the height $h(uP)+\Delta$ occurs along the branch that v terminates, and inserting uP as the parent of v and (if v was not the root node) the child of vP. The state cannot, however, be left like this as there is no guarantee that uP is still adjacent to a node in the same partition element as itself. So $\mathcal{P}$ is changed to a new partition $\mathcal{P}'$ as follows: if vP does not exist or v and vP are in the same element of $\mathcal{P}$, uP is moved to the element containing v. Otherwise, it is moved to either the element containing v or that containing vP with equal probability. This reallocation is enough to ensure that $\mathcal{P}'$ obeys connectedness rules. The effect on the transmission tree is that $d_{\mathcal{P}}(u)$ is moved to become a child of either $d_{\mathcal{P}}(v)$ or $d_{\mathcal{P}}(vP)$. If $d_{\mathcal{P}}(uS) \neq d_{\mathcal{P}}(uG)$ then $d_{\mathcal{P}}(uS)$ was the child of $d_{\mathcal{P}}(uG)$ before the move and remains so.

Noting that there are always $N-1$ choices for u, the Hastings ratio is the same as the standard subtree slide move, except that the denominator is multiplied by

$\frac{1}{2}$ if vP exists and v and vP are not in the same element of $\mathcal{P}$, and the numerator is multiplied by $\frac{1}{2}$ if uG exists and uG and uS are not in the same element of $\mathcal{P}$.

Type B Wilson-Balding move In a similar way, u is randomly picked from the set of nodes whose parents exist and are in different subtrees to themselves, and the standard Wilson-Balding move is performed on it, inserting uP as a parent of another node v and a child of its parent if that exists. The reassignment of uP to a new subtree is performed in the same was as for type B subtree slide, and the adjustment to the Hastings ratio is identical. The effect on the transmission tree is also the same.

5.3.3 Irreducibility of the chain

Suppose $\mathcal{P}$ is a partition of a phylogeny $\mathcal{G}$ with root node r and suppose $d_{\mathcal{P}}(r) = a_j$. We rely heavily on the fact that in the space of standard, unpartitioned phylogenies, the Wilson-Balding move on its own is sufficient for irreducibility [38]. Note that the following series of moves can transform a pair $\mathcal{G}, \mathcal{P}$ to a phylogeny in which, for each $a_i \in \mathbf{A}$, $c(a_i)$ and all its descendants are in the same partition element:

1. For any $a_i \in \mathbf{A}$, if $i \neq j$, a series of downward infection branch moves, starting with one on $e_{\mathcal{P}}(a_i)$, will eventually result in a partition $\mathcal{P}'$ in which $e_{\mathcal{P}'}(a_i) = c(a_i)$, in other words the earliest node u with $d_{\mathcal{P}}(u) = a_i$ is the most recent common ancestor of $d^{-1}(a_i)$.

2. As $c(a_i)$ now terminates the infection branch of a_i, use the type B Wilson-Balding move to make its parent $c(a_i)P$ the root node (if it is not already). After the move, $c(a_i)P$ will be in the same partition element as the old root node r.

3. Repeat this for all a_i with $i \neq j$.

Once this is completed, the result is a phylogeny and partition such that the only tips descended from $c(a_i)$ for each a_i (including a_j) are the members of the set $d^{-1}(a_i)$, and each $c(a_i)$ and all its descendants are in the same partition element. All nodes outside the clades rooted at each $c(a_i)$ are in the partition element containing $d^{-1}(a_j)$. In this tree, no host can root-block any other. From this partition and phylogeny, for any $k \neq j$, a sequence of upward infection branch moves, starting with one on $c(a_k)$ and going up to the child of the root that is its ancestor, followed by a sequence of downwards moves starting with the child of the root that is not $c(a_k)$'s ancestor and going down to the parent of $c(a_j)$, will change the partition only by reassigning all nodes outside the clades to the element containing $d^{-1}(a_k)$.

If $\mathcal{G}$, $\mathcal{P}$ and $\mathcal{G}'$, $\mathcal{P}'$ are any two phylogeny-partition pairs such that the tips corresponding to the same isolate have the same height in both trees, each may be transformed into a tree and partition of the above form such that all the nodes that are not descendants of any $c(a_i)$ are in the partition element that contains $d^{-1}(a_j)$ for an arbitrary $a_j \in \mathbf{A}$. A combination of type A and B Wilson-Balding moves and branch length changes can then be used to transform one tree and partition of this form to another, with the type A operator handling topological modifications within each clade (as, if all nodes are in the same partition, the type A version is simply the standard Wilson-Balding move) and the type B moving the clades. An example is shown in figure 5.5. This shows irreducibility as all these moves are reversible.

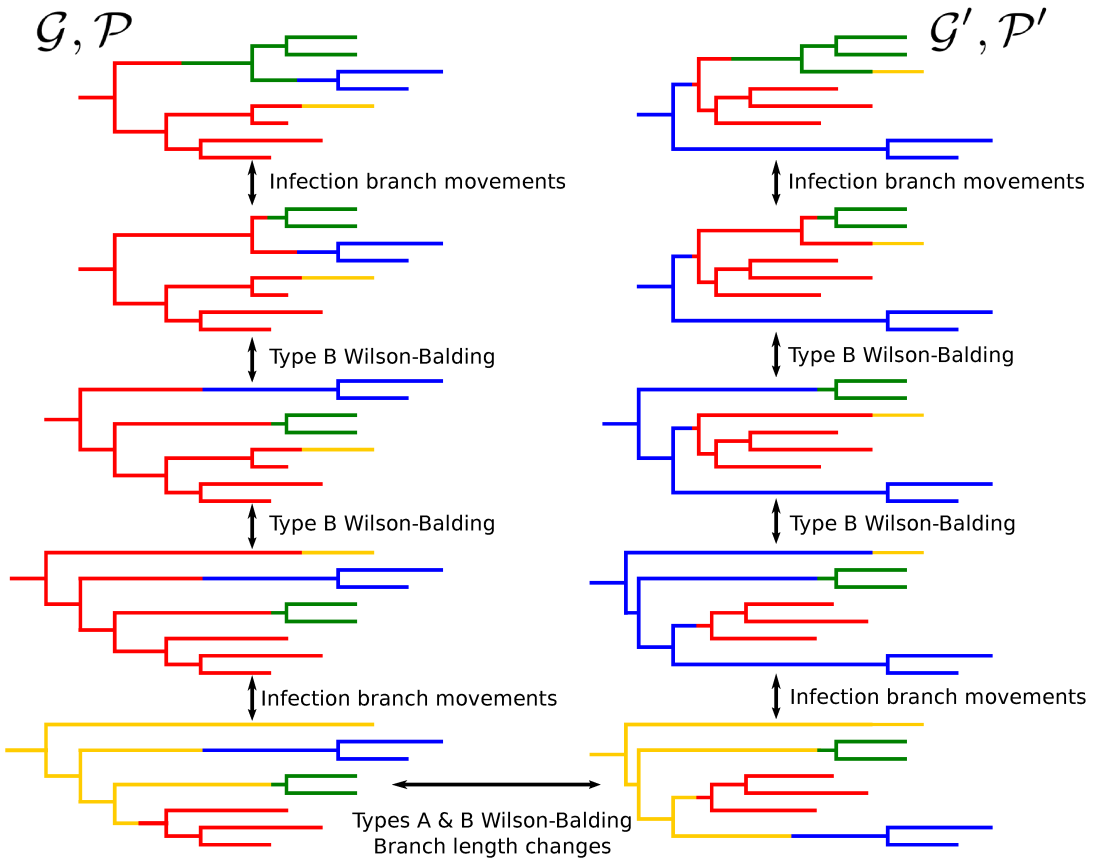


Figure 5.5: Illustration of the moves taking the phylogeny and partition $\mathcal{G}, \mathcal{P}$ (top left) to $\mathcal{G}', \mathcal{P}'$ (top right). Colours represent partition elements.

Chapter 6

Simultaneous exploration of the space of phylogenies and transmission trees: implementation

6.1 Introduction

The previous chapter outlined the theoretical framework behind treating transmission trees as partitions of the node set of a phylogeny, and outlined how this insight could be used to develop an MCMC framework to simultaneously reconstruct both types of tree. In this chapter I demonstrate how this can be done in practice. I give an example of a mathematical model of transmission and show how the posterior probability of a partitioned phylogeny given a set of sequence data can be calculated according to that model. I then apply this procedure, first to simulated datasets, and then to reanalyse the data from the Dutch H7N7 avian influenza outbreak of 2003. This outbreak has been the subject of many previous papers [15, 43, 140], including several that incorporated genetic data [10, 77, 148, 171,

172]. The paper by Ypma et al. [171], indeed, was one of the earliest methods for transmission tree reconstruction using genetic data, but was of the second type described in chapter 1, assuming no within-host genetic diversity and that coalescent events and transmission events coincide. The model in chapter 5 is of the third type, in which multiple independent lineages are allowed to coexist in the same host.

6.2 Methods

6.2.1 Major assumptions

The method here described is suitable for sequence data from pathogens infecting hosts in an outbreak or epidemic. In what follows, I make the following assumptions:

- As required by chapter 5, that transmission is a complete bottleneck and there is no reinfection or superinfection. Relaxation of the bottleneck assumption is not possible in the partitioned tree space as described in that chapter, but there may be ways in which reinfection or superinfection can be handled, at least in the absence of recombination or reassortment (see Discussion).

- Every infected host (or infected premises) in the outbreak has been identified, although it is not essential that each provides a sequence (see below). Relaxation of this assumption would require significant extra methodological work (see Discussion).

- Either a) the acquisition of a pathogen sequence from a host does not disturb the process of infection at all, or b) all hosts cease to be infectious once

a sequence is acquired. This could be relaxed if a different model of the infection process was used.

- The times of sampling for each sequence are known, and for any known hosts providing no sequence, a time at which infection was present is known. This cannot be relaxed.

- The noninfectiousness times of each host are known, or else they are known to have been infectious indefinitely. This is essential for this decomposition but alternatives exist in which it can be relaxed (for example, a construction similar to that used by Didelot et al. [35]).

- In the initial description I assume that hosts become infectious immediately after infection, but this is relaxed in a later section.

6.2.2 Model background

Suppose that the outbreak or epidemic infected N different hosts, and that during it, each of these N underwent one or more examinations which detected whether it was infected or not. Hosts that were found to be infected at an examination provided a pathogen isolate from which is obtained a nucleotide sequence (a positive examination); hosts that were not provided nothing (a negative examination). An examination could produce at most one sequence but multiple examinations could be performed simultaneously. Each host experienced at least one positive examination, so investigators are aware of all infections. The nucleotide sequences resulting from these examinations, together with information on negative examinations, forms the dataset D. It may be that there are known hosts in the epidemic for which no sequence is actually available; in these cases it is obvious that if an examination was made at a time at which it is known that infection was present, it would have been positive and provided a sequence, so I

insist that such an examination occurred but produced a noninformative sequence (i.e. consisting entirely of the code "N"). As a result there are M pathogen sequences with $N \leq M$. Denote the set of examination times by $\mathbf{T}^{\text{exam}}$. Hosts are ignored after their infection ends (due to presumed immunity, or the culling of infected animals) and subsequent examinations are discounted.

I describe the epidemiological and evolutionary processes involved in an epidemic with three models: a stochastic model of infection and between-host transmission dynamics, a deterministic model of the population dynamics of a within-host population of "agents", and a stochastic model of sequence evolution. Table 6.1 summarises the notation I will use to describe them. As in chapter 5, the transmission tree $\mathcal{N}$ is regarded as a map from each host to its infector.

In contrast to the previous work of Didelot et al. [35], whose underlying model of transmission was a compartmental SIR model, I use an individual-based model similar to those employed in previous work on agricultural outbreaks [26, 105, 171, 173]. This much more readily allows for the accommodation of host heterogeneity, and makes no assumption of random mixing. The process starts with a population of susceptible hosts. Some characteristics that allow us to define relationships between these hosts may be known *a priori*; if so, call these characteristics L. L could, for example, be the spatial locations of farms in an agricultural outbreak. The epidemic starts when a single susceptible is infected by an external source. If a_i is a host, t_i^{inf} is its time of infection. It is infectious from t_i^{inf} until a time t_i^{end}. The value of t_i^{end} is randomly determined at t_i^{inf}, by a draw from a probability distribution. Let $\mathbf{T}^{\text{inf}}$ be the complete set of infection times and $\mathbf{T}^{\text{end}}$ the complete set of noninfectiousness times. If a_i is infectious and a_j susceptible, a_i inflicts a constant force of infection on a_j given by a rate b modified by multiplication by a positive real number $F(a_i, a_j)$, where F is a positive function with parameters ϕ defining a relationship between a_i and a_j based on the information in L. In other words, the time between the infection of a_i and a possible infection of a_j

by a_i is drawn from an exponential distribution with mean $1/(bF(a_i, a_j))$. If the time drawn is such that a_i was no longer infectious at that point, or if some other infectious host had infected a_j at an earlier time, nothing happens. Otherwise, a_j becomes infected after this time. After t_i^{end}, a_i is considered removed and plays no further part in the epidemic.

There are many possible choices for F. If no spatial structure or other heterogeneity affecting transmission is assumed, then $F(a_i, a_j)$ can be set to 1 for all hosts a_i and a_j. Otherwise, it can be based on, for example, geographical distance between sampling sites, a network metric, or shared membership in some risk group. It can also be used to state prior information about the transmission tree structure; if it is known *a priori* that a_i did not infect a_j, then $F(a_i, a_j)$ can be set to zero. There is also no requirement that F be symmetric.

As outlined above, I assume that each host is examined at least once while it is infected, and in the more general case that examination does not disturb the course of the infection. Beyond that no concrete assumptions need to be made about the examination process; any number of examinations can be made of any hosts at any time. If examinations are instead restricted so that they only occur at at the point of noninfectiousness of each host, however, there are mathematical advantages, as will be seen.

As in previous work [35, 173] I take the model of the dynamics of the "agents" to be a coalescent process, with parameters ψ, amongst lineages in a freely-mixing population within each host. If the hosts are single organisms, the agents will naturally be individual pathogens. If, on the other hand, they are infected locations, they could instead be considered to be infected organisms. In either case, only a very small proportion of the total agent population are represented by lineages in the tree, and the assumption of a low sampling fraction required for use of the coalescent process is satisfied.

The sequence evolution model is of the standard type used in the reconstruction of time-resolved phylogenies [38]. It consists of both a continuous-time Markov chain model of sequence evolution (such as the commonly-used HKY [61] or GTR [120] models) and a molecular clock model. Denote the parameters of both by ω. I assume that mutation is a neutral process, and that it occurs independently of the host-to-host transmission structure.

Symbol	Type	Meaning
$\mathbf{T}^{\text{exam}}$	Background information	Examination times of each host
$\mathbf{T}^{\text{end}}$	Background information	Times of becoming noninfectious of each host
L	Background information	Information defining the relationship between hosts used to define F (e.g. spatial locations)
D	Data	Results of examinations (sequence data and notes of negative observations)
b	Model parameter	Unmodified transmission rate
ϕ	Model parameters	Parameters of F
ψ	Model parameters	Parameters of the population dynamics of the agents within each host
ω	Model parameters	Parameters of nucleotide substitution and molecular clock models
$\mathcal{G}$	Latent variable	Phylogenetic tree
$\mathcal{N}$	Latent variable	Transmission tree
$\mathbf{T}^{\text{inf}}$	Latent variables	Times of infection of each host
$\mathbf{T}^{\text{trans}}$	Latent variables	Times of infectiousness of each host (if different to $\mathbf{T}^{\text{inf}}$)
F	Function	Function modifying b based on known relationships between hosts

Table 6.1: Description of symbols used in the probability decomposition

6.2.3 Bayesian decomposition

In this section I show how the likelihood of a partitioned phylogeny can be calculated using the three models described above. We condition on $\mathbf{T}^{\text{exam}}$, $\mathbf{T}^{\text{end}}$, and L. If any or all hosts are known to have remained infectious indefinitely, the corresponding values of $\mathbf{T}^{\text{end}}$ can be set to the present day. It should be noted that the $\mathbf{T}^{\text{exam}}$ are not strictly sampling times. They instead represent times at which it is known that hosts were examined, and an infected host would provide a sequence. D is the results of these examinations, including the results of negative ones. This

formulation allows for some convenient mathematics but has consequences for estimation of the prior distribution (see section 6.4). Alternatively, if the data is such that all samples from each host were taken at the same time and the assumption that all hosts ceased to be infected immediately after this time is reasonable, $\mathbf{T}^{\text{exam}}$ need not be treated in this way and it is exactly analogous to the set of sampling times from other phylogenetic methods; D then consists solely of sequence data.

Ideally, it should be possible to enumerate all individuals or premises which were susceptible to infection but never experienced it, and L should include background information on them. For convenience we give these hosts infection times equal to the largest time in $\mathbf{T}^{\text{end}}$ (the time at which the last host became noninfectious) and undefined infectors and noninfectiousness times. This is necessary for unbiased estimation of b, which should only be interpreted as a transmission rate if such data is present in the analysis. If it is not, what is actually being estimated is a parameter b', which is what the unmodified transmission rate would be if all susceptibles did experience infection.

The posterior probability we are interested in calculating is $p(\mathcal{G}, \mathcal{N}, \mathbf{T}^{\text{inf}}, b, \phi, \psi, \omega | D, \mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L)$. By Bayes' Theorem this is equal to:

$$\frac{p(D|\mathcal{G}, \mathcal{N}, \mathbf{T}^{\text{inf}}, b, \phi, \psi, \omega, \mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L) p(\mathcal{G}, \mathcal{N}, \mathbf{T}^{\text{inf}}, b, \phi, \psi, \omega | \mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L)}{p(D|\mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L)}$$

As usual, the denominator need not be calculated if model comparison is not a consideration as it does not vary. If D may contain the results of negative examinations, it must be explicitly stated that if D includes M sequences but $\mathcal{G}$ has any number of tips other than M, then the probability of D given $\mathcal{G}$ is zero. A $\mathcal{G}$ with a different number of tips does not necessarily have zero prior probability, but it does result in zero likelihood for the data, so we need not concern ourself

with exploring the posterior probability space of such phylogenies. Given a $\mathcal{G}$ with the right number of tips, D depends by the assumptions of the mutation model only on $\mathcal{G}$ and ω, and the likelihood reduces to $p(D|\mathcal{G},\omega)$, which can be calculated using the Felsenstein pruning algorithm and the chosen molecular clock model in the normal way [37, 38, 46]. It remains to calculate the prior probability $p(\mathcal{G},\mathcal{N},\mathbf{T}^{\text{inf}},b,\phi,\psi,\omega|\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L)$. We decompose this as:

$$\begin{aligned} p(\mathcal{G},\mathcal{N},\mathbf{T}^{\text{inf}},b,\phi,\psi,\omega|\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) &= p(b|\mathcal{G},\mathcal{N},\mathbf{T}^{\text{inf}},\phi,\psi,\omega,\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) \\ &\quad \times p(\mathcal{G}|\mathcal{N},\mathbf{T}^{\text{inf}},\phi,\psi,\omega,\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) \\ &\quad \times p(\mathcal{N}|\mathbf{T}^{\text{inf}},\phi,\psi,\omega,\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) \\ &\quad \times p(\mathbf{T}^{\text{inf}}|\phi,\psi,\omega,\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) \\ &\quad \times p(\phi,\psi,\omega|\mathbf{T}^{\text{exam}},\mathbf{T}^{\text{end}},L) \end{aligned}$$

We make the following assumptions of conditional independence:

- All parameters are independent of ω; the mutation process has no bearing on the infection dynamics inside or between hosts.

- The base transmission rate b is conditionally independent of $\mathcal{G}$, $\mathbf{T}^{\text{exam}}$, and ψ given ϕ, $\mathcal{N}$, $\mathbf{T}^{\text{inf}}$, $\mathbf{T}^{\text{end}}$, and L. It can be determined if the transmission tree, timings of the epidemic, and other parameters of the between-host model are known and we assume it is not affected by examination.

- The phylogeny $\mathcal{G}$ is conditionally independent of ϕ, $\mathbf{T}^{\text{end}}$, and L given ψ, $\mathcal{N}$, $\mathbf{T}^{\text{inf}}$, and $\mathbf{T}^{\text{exam}}$. L and ϕ determine the transmission model and are not relevant if the full transmission tree and its timings are already known, whereas the tips of the phylogeny correspond to $\mathbf{T}^{\text{exam}}$, not $\mathbf{T}^{\text{end}}$.

- The transmission tree $\mathcal{N}$ is conditionally independent of $\mathbf{T}^{\text{exam}}$ and ψ given ϕ, $\mathbf{T}^{\text{inf}}$, $\mathbf{T}^{\text{end}}$ and L. Once again, parameters of the within-host model are

not relevant to the between-host model and examination is assumed to not disturb the transmission tree.

- The infection times $\mathbf{T}^{\text{inf}}$ are conditionally independent of ϕ, $\mathbf{T}^{\text{exam}}$, ψ and L given $\mathbf{T}^{\text{end}}$. I assume the infected period of each host is unaffected by its relationship with other hosts, and the formulation above is such that a host's actual infection status at the time of examination is part of D, not $\mathbf{T}^{\text{exam}}$.

- ϕ, ψ and ω are independent of $\mathbf{T}^{\text{exam}}$, $\mathbf{T}^{\text{end}}$ and each other. The parameters determining transmission, within-host growth, and mutation are independent of each other, of the examination process, of the times of noninfectiousness of this particular epidemic, and of the exact relationships amongst this set of hosts.

The decomposition is therefore reduced to:

$$\begin{aligned} p(\mathcal{G}, \mathcal{N}, \mathbf{T}^{\text{inf}}, b, \phi, \psi, \omega | \mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L) &= p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \phi, \mathbf{T}^{\text{end}}, L) \\ &\quad \times p(\mathcal{G}|\mathcal{N}, \mathbf{T}^{\text{inf}}, \psi, \mathbf{T}^{\text{exam}}) \\ &\quad \times p(\mathcal{N}|\mathbf{T}^{\text{inf}}, \phi, \mathbf{T}^{\text{end}}, L) \\ &\quad \times p(\mathbf{T}^{\text{inf}}|\mathbf{T}^{\text{end}}) \\ &\quad \times p(\phi)p(\psi)p(\omega) \end{aligned}$$

For calculation of $p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \phi, \mathbf{T}^{\text{end}}, L)$, we use Bayes' Theorem again:

$$p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \phi, \mathbf{T}^{\text{end}}, L) = \frac{p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, L)p(b|\phi, L)}{p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|\phi, L)} \tag{6.1}$$

and note that the denominator can be evaluated as

$$\int_0^\infty p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, L)p(b|\phi, L)db$$

by the law of total probability.

The term $p(b|\phi, L)$ is the prior belief in the value of b given ϕ and background information; we take b to be independent of these and let $p(b)$ be the improper uniform distribution on $[0, \infty)$ or a gamma distribution. The term $p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, L)$ is the probability that the observed transmission tree and all its timings occurred for a given b and ϕ. This can be calculated using a procedure similar to that employed by Deardon et al. [32]. If there are, in addition to the N infected hosts $a_1, \ldots, a_N$, N' known potential hosts $a_{N+1}, \ldots, a_{N+N'}$ that were never infected, let o be a permutation function such that $t^{\text{inf}}_{o(1)}, \ldots, t^{\text{inf}}_{o(N+N')}$ is in increasing order of time (breaking ties in an arbitrary, but deterministic, fashion and remembering that never-infected hosts are given infection times after those of any infected hosts).

Each $a_i \in \mathbf{A}$ is infected at t^{inf}_i and ceases to be infectious at t^{end}_i. Suppose that we have a set of parameters ρ that define random variables such that, for each $a_i \in \mathbf{A}$, we can write down a probability $p(t^{\text{end}}_i|t^{\text{inf}}_i, \rho)$. We need not make this function explicit for reasons that will become clear. Elements of ρ may be dependent on each other and background information in complex ways, but crucially each t^{end}_i is conditionally independent of b, ϕ, L, o, $\mathcal{N}$ and t^{inf}_j for $j \neq i$ given t^{inf}_i, and ρ.

Note that knowledge of $\mathbf{T}^{\text{inf}}$ combined with an arbitrary deterministic means of breaking ties defines o, so we can write $p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, \rho, L) = p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}, o|b, \phi, \rho, L)$. We decompose as follows:

$$\begin{aligned} &p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}, o|b, \phi, \rho, L) \\ &\quad = \prod_{i=1}^{N+N'} p(\mathcal{N}(a_{o(i)}), t^{\text{inf}}_{o(i)}, t^{\text{end}}_{o(i)}, o(i)|\{\mathcal{N}(a_{o(j)}), t^{\text{inf}}_{o(j)}, t^{\text{end}}_{o(j)}, o(j) : j < i\}, b, \phi, \rho, L) \end{aligned} \tag{6.2}$$

and consider each term in the product individually. There are three possibilities.

Case 1: $i > N$**.** In this case $a_{o(i)}$ was never infected, and this combination of $\mathcal{N}(a_{o(i)})$, $t^{\text{inf}}_{o(i)}$, $t^{\text{end}}_{o(i)}$ and $o(i)$ represents this state of affairs. If o and the time $t^{\text{inf}}_{o(N)}$

of the last infection are known, then this is the probability no members of $\mathcal{I}(t^{\text{inf}}_{o(N)})$ infected $a_{o(i)}$. This is equal to

$$W_i = \prod_{a_j \in \mathcal{I}(t^{\text{inf}}_{o(N)})} \left(-bF(a_j, a_i)(t^{\text{end}}_j - t^{\text{inf}}_N)\right)$$

Case 2: $2 \leq i \leq N$**.** Firstly, as outlined above we assume we can write:

$$\begin{aligned}
&p(\mathcal{N}(a_{o(i)}), t^{\text{inf}}_{o(i)}, t^{\text{end}}_{o(i)}, o(i) | \{\mathcal{N}(a_{o(j)}), t^{\text{inf}}_{o(j)}, t^{\text{end}}_{o(j)}, o(j) : j < i\}, b, \phi, \rho, L) \\
&= p(t^{\text{end}}_{o(i)} | t^{\text{inf}}_{o(i)}, \rho) p(\mathcal{N}(a_{o(i)}), t^{\text{inf}}_{o(i)}, o(i) | \{\mathcal{N}(a_{o(j)}), t^{\text{inf}}_{o(j)}, t^{\text{end}}_{o(j)}, o(j) : j < i\}, b, \phi, \rho, L)
\end{aligned}$$

Knowledge of $o(j)$ for $j < i$ implies that all $a_{o(k)}$ with $k \geq i$ are members of $\mathcal{S}(T^{\text{inf}}_{o(i-1)})$. The term $p(\mathcal{N}(a_{o(i)}), t^{\text{inf}}_{o(i)}, o(i) | \{\mathcal{N}(a_{o(j)}), t^{\text{inf}}_{o(j)}, t^{\text{end}}_{o(j)}, o(j) : j < i\}, b, \phi, \rho, L)$ is then the probability that $a_{o(i)}$ is the next case to be infected, and that it is infected by $\mathcal{N}(a_{o(i)})$ at $t^{\text{inf}}_{o(i)}$. This is the product of three individual terms:

- X_i, the probability that $\mathcal{N}(a_{o(i)})$ infected $a_{o(i)}$ at $t^{\text{inf}}_{o(i)}$, but not at any time before that during the interval $[t^{\text{inf}}_{o(i-1)}, t^{\text{inf}}_{o(i)}]$:

$$X_i = bF(a_{o(i)}, \mathcal{N}(a_{o(i)})) \times \exp\left(-bF(a_{o(i)}, \mathcal{N}(a_{o(i)}))(t^{\text{inf}}_{o(i)} - t^{\text{inf}}_{o(i-1)})\right)$$

- Y_i, the probability that no host in $\mathcal{I}(t^{\text{inf}}_{o(i-1)})$ other than $\mathcal{N}(a_{o(i)})$ infected $a_{o(i)}$ before $t^{\text{inf}}_{o(i)}$ in that interval. Noting that the upper bound on the time that such an a_j could have infected $a_{o(i)}$ before $t^{\text{inf}}_{o(i)}$ is either $t^{\text{inf}}_{o(i)}$ itself if a_j was still infectious at that point or t^{end}_j if it was not, this is given by:

$$Y_i = \prod_{\substack{a_j \in \mathcal{I}(t^{\text{inf}}_{o(i-1)}) \\ a_j \neq \mathcal{N}(a_{o(i)})}} \exp\left(-bF(a_{o(i)}, a_j)(\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_j\} - t^{\text{inf}}_{o(i-1)})\right)$$

- Z_i, the probability that no host in $\mathcal{I}(t^{\text{inf}}_{o(i-1)})$ infected any host other than

$a_{o(i)}$ in $\mathcal{S}(t^{\text{inf}}_{o(i-1)})$ during that interval. Again, the upper bound on the time at which an a_j could infect a third host a_k before $t^{\text{inf}}_{o(i)}$ is $\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_j\}$.

$$Z_i = \prod_{a_j \in \mathcal{I}(t^{\text{inf}}_{o(i-1)})} \prod_{\substack{a_k \in \mathcal{S}(t^{\text{inf}}_{o(i-1)}) \\ k \neq o(i)}} \exp\left(-bF(a_j, a_k)(\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_j\} - t^{\text{inf}}_{o(i-1)})\right)$$

Case 3: $i = 1$**.** The term is $p(\mathcal{N}(a_{o(1)}), t^{\text{inf}}_{o(1)}, t^{\text{end}}_{o(1)}, o(1)|b, \phi, \rho, L)$ which can be written as the product $p(t^{\text{end}}_{o(1)}, |t^{\text{inf}}_{o(1)}, \rho)p(\mathcal{N}(a_{o(1)}), t^{\text{inf}}_{o(1)}, o(1)|b, \phi, L)$. The second half of this product is the probability of the index infection, which is effectively unknowable and we set to 1.

Returning to (6.2) we have:

$$p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, \rho, L) = \prod_{i=N+1}^{N+N'} W_i \prod_{i=2}^{N} X_i Y_i Z_i \prod_{i=1}^{N} p(t^{\text{end}}_{o(i)}|t^{\text{inf}}_{o(i)}, \rho) \tag{6.3}$$

We then do some regrouping. The terms relevant to the infectious pressure exerted by $a_{o(j)}$ on $a_{o(i)}$, prior to the infection of the latter at $t^{\text{inf}}_{o(i)}$ (whether this represents a real infection or not), where $j < i$ and l is the (unique) index such that $a_{o(j)} \in \mathcal{I}(t^{\text{inf}}_{o(l-1)})$ but $a_{o(j)} \notin \mathcal{I}(t^{\text{inf}}_{o(l)})$, are:

$$\prod_{k \in \{j+1, \ldots, \min\{l,i\}\}} \exp\left(-bF(a_{o(i)}, a_{o(j)})(\min\{t^{\text{inf}}_{o(k)}, t^{\text{end}}_{o(j)}\} - t^{\text{inf}}_{o(k-1)})\right)$$
$$= \exp\left(-bF(a_{o(i)}, a_{o(j)})(\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_{o(j)}\} - t^{\text{inf}}_{o(j)})\right)$$

and if we regroup all of (6.3) this way, it becomes:

$$p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, \rho, L) = Db^{N-1}\exp(-bE) \prod_{i \in \{1, \ldots, N\}} p(t^{\text{end}}_i|t^{\text{inf}}_i, \rho)$$

where

$$D = \prod_{i\in\{2,\ldots,N\}} F(a_{o(i)}, \mathcal{N}(a_{o(i)}))$$

and

$$E = \sum_{i\in\{2,\ldots,N\}} \sum_{j\in\{1,\ldots,i-1\}} F(a_{o(i)}, a_{o(j)})(\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_{o(j)}\} - t^{\text{inf}}_{o(j)}) + \sum_{i\in\{N+1\ldots,N+N'\}} \sum_{j\in\{1,\ldots,N\}} F(a_{o(i)}, a_{o(j)})(t^{\text{end}}_{o(j)} - t^{\text{inf}}_{o(j)}) \quad (6.4)$$

We marginalise out ρ to get

$$\begin{aligned} p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, L) &= \int_\rho p(\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}|b, \phi, \rho, L)p(\rho)d\rho \\ &= D\left(\int_\rho \prod_{i\in\{1,\ldots,N\}} p(t^{\text{end}}_i|t^{\text{inf}}_i, \rho)p(\rho)d\rho\right) b^{N-1}\exp(-bE) \end{aligned}$$

Returning to (6.1), the denominator can be rearranged:

$$\begin{aligned} p(\mathcal{N},\mathbf{T}^{\text{inf}},\mathbf{T}^{\text{end}}|\phi,L) &= \int_b p(\mathcal{N},\mathbf{T}^{\text{inf}},\mathbf{T}^{\text{end}}|b,\phi,L)p(b)db \\ &= \int_0^\infty D\left(\int_\rho \prod_{i\in\{1,\ldots,N\}} p(t^{\text{end}}_i|t^{\text{inf}}_i,\rho)p(\rho)d\rho\right) b^{N-1}\exp(-bE)p(b)db \\ &= D\int_\rho \prod_{k\in\{1,\ldots,N\}} p(t^{\text{end}}_k|t^{\text{inf}}_k,\rho)p(\rho)d\rho \int_0^\infty b^{N-1}\exp(-bE)p(b)db \end{aligned}$$

When (6.1) is formed, both D and the integral involving ρ cancel, to get:

$$p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}, \phi, L) = \frac{b^{N-1}\exp(-bE)p(b)}{\int_0^\infty b^{N-1}\exp(-bE)p(b)db}$$

Suppose first that $p(b) = 1$; i.e. the prior on b is uniform improper. As

$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}}$, we have that:

$$p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}, \phi, L) = \frac{E^N b^{N-1} \exp(-bE)}{\Gamma(N)}$$

which is in fact exactly the same as saying that b is gamma distributed with shape N and rate E. Alternatively, if we make the the prior on b a gamma distribution with shape x and rate y, then:

$$p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{end}}, \phi, L) = \frac{(E+y)^{N+x-1} b^{N+x-2} \exp(-b(E+y))}{\Gamma(N+x-1)}$$

or b is gamma distributed with shape $N + x - 1$ and rate $E + y$.

If we do not have information on uninfected susceptibles and are therefore estimating b' rather than b, then only the first part of the sum in (6.4) is present. If we call this E' then clearly $E' < E$ and as a result the expected value of b' is greater than that of b.

Next, we need to calculate $p(\mathcal{G}|\mathcal{N}, \mathbf{T}^{\text{inf}}, \psi, \mathbf{T}^{\text{exam}})$. The first observation we make is that, since examinations never take place after noninfectiousness, the combination of $\mathbf{T}^{\text{inf}}$ and $\mathbf{T}^{\text{exam}}$ determines which examinations were positive, and that positive examinations correspond to the tips in $\mathcal{G}$. If the number of positive examinations of a given host and the number of tips in $\mathcal{G}$ corresponding to sequences taken from that host differ, then this term must be equal to zero. In theory, we can calculate this term for a phylogeny with any number of tips up to the total number of examinations, but in practice we need not if we are sampling from the posterior distribution, as any tree that does not have M tips will have zero posterior probability because the likelihood will be zero. So we can assume that $\mathcal{G}$ has M tips and that no tip date is before the infection date of the corresponding host, and merely check that $\mathbf{T}^{\text{inf}}$ and $\mathbf{T}^{\text{exam}}$ imply M positive observations.

If the tip count is correct, we then calculate this probability by extending the

procedure outlined by Didelot et al. [35] to allow for the use of any of the standard models of deterministic population growth, and the possibility of host heterogeneity. The latter is accomplished by dividing the set of hosts into categories and assigning a separate demographic model to all of those in each one. Categories can be assigned from known epidemiological data about the hosts; for example, in a livestock disease outbreak, they may reflect the size of farm. Naturally, there is no requirement that there be more than one category. If $\mathbf{c}$ is such a category, there is a corresponding demographic function $N_{\mathbf{c}} : \mathbb{R}$ with parameters $\psi_{\mathbf{c}}$ where $N_{\mathbf{c}}(t)$ is the product of the effective population size and generation time of the agents at time t on a separate backwards timescale in each host. Let $cc(i)$ be the category that a_i belongs to.

Suppose that, according to $\mathcal{N}$, $\mathbf{T}^{\text{inf}}$ and $\mathbf{T}^{\text{exam}}$, $a_i \in \mathbf{A}$ infected n_i other hosts and that there were m_i positive observations of a_i. Suppose $\mathcal{H}_i$ is a phylogenetic tree that describes the part of the outbreak that took place within a_i. Because we assume transmission is a complete bottleneck, it is a single tree with a root note r. It will have $n_i + m_i$ tips, one for each infection event and each positive observation. If the time of the root r is t_i^{root}, we know that t_i^{root} is later than t_i^{inf} and we give $\mathcal{H}_i$ a root branch of length $t_i^{\text{root}} - t_i^{\text{inf}}$. If we have a $\mathcal{H}_i$ for each a_i, and we know $\mathcal{N}$, we can build a phylogenetic tree for the entire epidemic by, if $a_j = \mathcal{N}(a_i)$, attaching the root node of $\mathcal{H}_i$ to the tip of $\mathcal{H}_j$ that corresponds to the infection of a_i by a branch with length equal to the root branch length of $\mathcal{H}_i$. If $\mathcal{G}$ cannot be built up from $\mathcal{H}_i$s in this way, $p(\mathcal{G}|\mathcal{N}, \mathbf{T}^{\text{inf}}, \psi, \mathbf{T}^{\text{exam}}) = 0$. Otherwise, we calculate it as:

$$p(\mathcal{G}|\mathcal{N}, \mathbf{T}^{\text{inf}}, \psi, \mathbf{T}^{\text{exam}}) = \prod_{i \in \{1,\ldots,N\}} p(\mathcal{H}_i|\psi_{cc(i)})$$

In the standard coalescent model [134], the probability density function for the time t (in backwards time) of the first coalescence of $K \geq 2$ lineages after t_0 where

the demographic function is $N_{\mathbf{c}}$ is given by:

$$p(t) = \frac{K(K-1)}{2N_{\mathbf{c}}(t)} \exp\left(-\int_{t_0}^{t} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right)$$

and if we know which two specific lineages coalesced, the first $K(K-1)/2$ cancels. As Didelot et al. [35] note, this is not quite sufficient for our purposes because we have a maximum height for the last coalescence. If this is $t_{\max}$, the normalised probability distribution for the time of first coalescence is:

$$p(t|t_{\max}) = \begin{cases} \frac{\frac{K(K-1)}{2N_{\mathbf{c}}(t)} \exp\left(-\int_{t_0}^{t} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right)}{1-\exp\left(-\int_{t_0}^{t_{\max}} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right)} & t_0 \leq t < t_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

This is the probability of an interval in $\mathcal{H}_i$ ending in a coalescent event. The probability of an interval ending in a transmission or sampling event is the probability that no events occur in the interval, which is one minus the cumulative distribution function $P(t|t_{\max})$:

$$1 - P(t|t_{\max}) = \begin{cases} 1 & t < t_0 \\ \frac{\exp\left(-\int_{t_0}^{t} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right) - \exp\left(-\int_{t_0}^{t_{\max}} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right)}{1-\exp\left(-\int_{t_0}^{t_{\max}} \frac{K(K-1)}{2N_{\mathbf{c}}(s)} ds\right)} & t_0 \leq t < t_{\max} \\ 0 & t \geq t_{\max} \end{cases} \tag{6.6}$$

Note that while with no maximum root height, the formula happens to work for $K = 1$, here it does not as the denominator is 0 for $t_0 \leq t < t_{\max}$, and we instead set the probability of any interval with one lineage to 1. In particular, if a_i has no children then $p(\mathcal{H}_i|\psi_{cc(i)}) = 1$.

These formulae can be used to calculate $p(\mathcal{H}_i|\psi_{cc(i)})$ for every $\mathcal{H}_i$ in the established way for a tree with temporally offset tips [38]. It is most intuitive to standardise the timescale of each $\mathcal{H}_i$ such that the effective population size at the point of the infection can be the same across all hosts. As a result, I depart from the

convention of making height 0 the time of the last tip, and instead put it at the time of infection (i.e. $t_{\max} = 0$), with all later events occurring at negative heights. Appropriate demographic functions should be picked for the $N_{\mathbf{c}}$s; I suggest exponential or logistic growth [114, 134].

The next term in the decomposition is $p(\mathcal{N}|\mathbf{T}^{\inf}, \phi, \mathbf{T}^{\text{end}}, L)$. Recall that $\mathcal{I}(t)$ is the set of infected hosts at time t. The instantaneous probability that host a_i was infected by host $\mathcal{N}(a_i)$ at time $t_i^{\inf}$ is $bF(a_i, \mathcal{N}(a_i))\mathbb{1}_{a_i \in \mathcal{I}(t_i^{\inf})}$, and if we condition on the fact that a_i was indeed infected by *some* host at $t_i^{\inf}$ then we normalise by $\sum_{a_j \in \mathcal{I}(t_i^{\inf})} bF(a_i, a_j)$. In this normalisation the bs cancel. The probability of the infection of the first host in the epidemic from its external source is set to 1. The expression is:

$$p(\mathcal{N}|\mathbf{T}^{\inf}, \phi, \mathbf{T}^{\text{end}}, L) = \prod_{\substack{a_i \in \mathbf{A} \\ \mathcal{N}(a_i) \neq \emptyset}} \frac{F(a_i, \mathcal{N}(a_i))\mathbb{1}_{a_j \in \mathcal{I}(t_i^{\inf})}}{\sum_{a_j \in \mathcal{I}(t_i^{\inf})} F(a_i, a_j)}$$

The calculation of $p(\mathbf{T}^{\inf}|\mathbf{T}^{\text{end}})$, the probability of the times of infection, can be handled in a number of ways. These terms can be seen as the probabilites of the times from infection to noninfectiousness, $t_i^{\text{end}} - t_i^{\inf}$, of each a_i. Previous work on foot-and-mouth disease virus [26, 105] has used clinical data to estimate times of infection, and if this kind of information is available, it can be used to determine a separate prior distribution for each $t_i^{\text{end}} - t_i^{\inf}$. This is also the preferable approach if infections are ongoing at the time of sampling. If information of this type is not available, a similar approach to that in the coalescent calculations above can be taken, assigning each host a_i to an infectious period category $ic(a_i)$. This again allows the procedure to accommodate known heterogeneity; for example in an agricultural outbreak it is likely that infectious periods decrease as time goes by and control measures are brought to bear.

If the infectious period of the disease is well understood, a single prior distribution

for $t_i^{\text{end}} - t_i^{\text{inf}}$ can be assigned for all hosts in each category. It may be, however, that a user would hope to estimate the distribution of infectious periods from the genetic data. In this case, each $t_i^{\text{end}} - t_i^{\text{inf}}$ within a category can be taken as a draw from a probability distribution with unknown parameters, and then hyperpriors can be put on those parameters. Rather than using MCMC to estimate both the parameters χ of a probability distribution D and a set of draws from that distribution, I choose to integrate out the actual values of χ by using D's conjugate prior for them and then calculating the marginal likelihood of the infectious periods given the hyperpriors. Any continuous probability distribution can be considered if this marginal likelihood is analytically tractable. Examples are normal, lognormal, exponential, and gamma if the shape parameter is known. Although it it not absolutely ideal as infectious periods are non-negative parameters, I suggest the normal distribution as the prior for the reason that its mean and variance are independent.

Finally, all that remains is to place prior distributions on the parameters making up ϕ, ψ, and ω.

6.2.4 Latent periods

The above formulation has taken the course of infection to follow a SIR process; hosts are assumed to be infectious as soon as they are infected. It is straightforward to replace this with a SEIR process instead. While it is possible to handle latent periods using a hyperprior in the same way as described for infectious periods in the previous section, in simulations this resulted in poor mixing of the MCMC chain if an strongly informative hyperprior on the distribution of their lengths was used, and poor estimation of their values if the hyperprior was weaker. Instead, we again subdivide the set of hosts into discrete categories and assign a single value to the latent period for all hosts in each category, so that the latent period of

host a_i is $lc(a_i)$. Let $\mathbf{T}^{\text{trans}}$ be the set of infectiousness times of each host; then if $t_i^{\text{trans}} \in \mathbf{T}^{\text{trans}}$ is the infectiousness time of a_i, $t_i^{\text{trans}} = t_i^{\text{inf}} + lc(a_i)$. We assume that hosts are infectious by the time they cease to be infected, and that examinations of infected but noninfectious hosts are positive. The phylogeny $\mathcal{G}$ is assumed to be conditionally independent of $\mathbf{T}^{\text{trans}}$ given $\mathbf{T}^{\text{inf}}$, and $\mathbf{T}^{\text{inf}}$ of $\mathbf{T}^{\text{end}}$ given $\mathbf{T}^{\text{trans}}$.

The new decomposition is:

$$\begin{aligned} p(\mathcal{G}, \mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{trans}}, \phi, \psi, \omega | \mathbf{T}^{\text{exam}}, \mathbf{T}^{\text{end}}, L) &= p(b|\mathcal{N}, \mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{trans}}, \phi, \mathbf{T}^{\text{end}}, L) \\ &\quad \times p(\mathcal{G}|\mathcal{N}, \mathbf{T}^{\text{inf}}, \psi, \mathbf{T}^{\text{exam}}) \\ &\quad \times p(\mathcal{N}|\mathbf{T}^{\text{inf}}, \mathbf{T}^{\text{trans}}, \phi, \mathbf{T}^{\text{end}}, L) \\ &\quad \times p(\mathbf{T}^{\text{inf}}|\mathbf{T}^{\text{trans}}) \\ &\quad \times p(\mathbf{T}^{\text{trans}}|\mathbf{T}^{\text{end}}) \\ &\quad \times p(\phi)p(\psi)p(\omega) \end{aligned}$$

It is not a major modification to the SIR version to calculate the first term in this product; the key difference is that a new version of (6.2) is required. We now assume that for any a_i, both the time of infectiousness t_i^{trans} and t_i^{end} are determined at t_i^{inf}. If ρ now contains information on the distribution of latent periods as well as infectious periods, we suppose $p(t_i^{\text{end}}, t_i^{\text{trans}}|t_i^{\text{inf}}, \rho)$ can be written down but once again it will cancel out in the normalisation. Non-infected hosts are given undefined infectiousness times as well as noninfectiousness times. We introduce a set $\mathcal{E}(t)$ of those hosts which were infected but not infectious at time t. Once again, let o sort the hosts into increasing order of infection. For notational

convenience, write $\mathbf{n}(i)$ for the index of the infector of a_i.

$$\begin{aligned}
&p(\mathcal{N},\mathbf{T}^{\text{inf}},\mathbf{T}^{\text{trans}},\mathbf{T}^{\text{end}},o|b,\phi,\rho,L)\\
&=\prod_{i=1}^{N+N'} p(\mathcal{N}(a_{o(i)}),t^{\text{inf}}_{o(i)},t^{\text{trans}}_{o(i)},t^{\text{end}}_{o(i)},o(i)|\{\mathcal{N}(a_{o(j)}),t^{\text{inf}}_{o(j)},t^{\text{trans}}_{o(i)},t^{\text{end}}_{o(j)},o(j){:}j{<}i\},b,\phi,\rho,L)
\end{aligned} \tag{6.7}$$

Once again, we calculate the product term-by-term. Let $\mathcal{EI}(t) = \mathcal{E}(t) \cup \mathcal{I}(t)$. Then $\mathcal{EI}(t_1) \cap \mathcal{I}(t_2)$, if $t_1 < t_2$, is the set of hosts that were infected before t_1 and infectious by t_2.

Case 1: $i > N$**.** W_i is modified as follows:

$$W_i \;=\; \prod_{a_j \in \mathcal{EI}(t^{\text{inf}}_{o(i-1)}) \cap \mathcal{I}(t^{\text{inf}}_{o(i)})} \left(-bF(a_j,a_i)(t^{\text{end}}_j - \max\{t^{\text{inf}}_{o(N)}, t^{\text{trans}}_j\})\right)$$

Case 2: $2 \le i \le N$**.** If latent and infectious periods depends only on ρ, the term can be written as follows:

$$p(t^{\text{trans}}_{o(i)},t^{\text{end}}_{o(i)}|t^{\text{inf}}_{o(i)},\rho)p(\mathcal{N}(a_{o(i)}),t^{\text{inf}}_{o(i)},o(i)|\{\mathcal{N}(a_{o(j)}),t^{\text{inf}}_{o(j)},t^{\text{trans}}_{o(j)},t^{\text{end}}_{o(j)},o(j){:}j{<}i\},b,\phi,\rho,L)$$

The second term of this product is composed of slightly different versions of X_i, Y_i and Z_i, accounting for the fact that membership of $\mathcal{I}(t^{\text{inf}}_{o(i-1)})$ cannot now be assumed for cases that were infected at $t^{\text{inf}}_{o(i-1)}$.

$$\begin{aligned}
X_i &= bF(a_{o(i)},a_{\mathbf{n}(o(i))}) \times \exp\Big(-bF(a_{o(i)},a_{\mathbf{n}(o(i))})(t^{\text{inf}}_{o(i)} - \max\{t^{\text{inf}}_{o(i-1)},t^{\text{trans}}_{\mathbf{n}(o(i))}\})\Big)\\
Y_i &= \prod_{\substack{a_j \in \mathcal{EI}(t^{\text{inf}}_{o(i-1)}) \cap \mathcal{I}(t^{\text{inf}}_{o(i)})\\ a_j \neq \mathcal{N}(a_{o(i)})}} \exp\Big(-bF(a_{o(i)},a_j)(\min\{t^{\text{inf}}_{o(i)},t^{\text{end}}_j\} - \max\{t^{\text{inf}}_{o(i-1)},t^{\text{trans}}_j\})\Big)\\
Z_i &= \prod_{\substack{a_j \in \mathcal{EI}(t^{\text{inf}}_{o(i-1)})\\ \cap \mathcal{I}(t^{\text{inf}}_{o(i)})}} \left(\prod_{\substack{a_k \in \mathcal{S}(t^{\text{inf}}_{o(i-1)})\\ k \neq o(i)}} \exp\Big(-bF(a_j,a_k)(\min\{t^{\text{inf}}_{o(i)},t^{\text{end}}_j\} - \max\{t^{\text{inf}}_{o(i-1)},t^{\text{trans}}_j\})\Big) \right)
\end{aligned}$$

Case 3: $i = N$**.** The probability of the index infection is assumed to be 1 as before, so this term is simply $p(t^{\text{trans}}_{o(1)}, t^{\text{end}}_{o(1)} | t^{\text{inf}}_{o(1)}, \rho)$.

We form the product (6.7) and regroup as before. D is unchanged, and E requires only the modification that infectiousness no longer starts at the point of infection:

$$E = \sum_{i \in \{2,\ldots,N\}} \sum_{j \in \{1,\ldots,i-1\}} F(a_{o(i)}, a_{o(j)})(\min\{t^{\text{inf}}_{o(i)}, t^{\text{end}}_{o(j)}\} - t^{\text{trans}}_{o(j)}) + \sum_{i \in \{N+1\ldots,N+N'\}} \sum_{j \in \{1,\ldots,N\}} F(a_{o(i)}, a_{o(j)})(t^{\text{end}}_{o(j)} - t^{\text{trans}}_{o(j)})$$

The rest of the derivation is identical to that of the previous section.

Calculating the third term of the full decomposition merely involves accounting for the fact that infectious pressure is now only applied after the end of a host's latent period when determining $\mathbb{1}_{a_i \in \mathcal{I}(t^{\text{inf}}_i)}$. The term $p(\mathbf{T}^{\text{inf}} | \mathbf{T}^{\text{trans}})$ is based on assigning a prior distribution to the latent periods of every category, while $p(\mathbf{T}^{\text{trans}} | \mathbf{T}^{\text{end}})$ simply replaces $p(\mathbf{T}^{\text{inf}} | \mathbf{T}^{\text{end}})$ in the original decomposition as infectious periods no longer start at infection.

6.2.5 Simulations

Epidemics and sequences were simulated using examples of the three models described above. The simulations were intended to represented a situation analogous to an agricultural outbreak, with the hosts as farms. The units of time were intended to represent days. In each replicate of the simulation, $\mathbf{A}$ consisted of 50 potential hosts arranged spatially on a regular 5×5 grid contained in the unit square, such that every grid point contained two whose distance from each other was zero. A single host was chosen at random to be infected first. The infection of each followed a SEIR process: upon infection, a host a_i was latently

infected for a time P^{lat} which was identical across all hosts and subsequently infectious for a period p_i^{inf} drawn from a normal distribution (negative draws were discarded, but the distribution used was such that the probability of these occurring was negligible). Let $\mathbf{P}^{\text{inf}}$ be the set of all the infectious periods.

F was an exponential spatial transmission kernel function: the time for an newly infectious a_i to infect a susceptible a_j was drawn from an exponential distribution with mean $be^{-\alpha d(a_i,a_j)}$ where $d(a_i, a_j)$ is the Euclidean distance between the locations of a_i and a_j. The process was run until no infections remained. A single positive examination was simulated at the point of noninfectiousness of each host. As no infections persisted following the acquisition of a sequence, this is the special case outlined above and so there is no need to consider the possibility of negative examinations in the analysis. Only simulations in which at least 45 of the 50 susceptibles were eventually infected were kept.

Once the epidemic simulation was completed, the transmission tree was transformed into a phylogenetic tree by simulating a within-host phylogeny under a coalescent process. Variation in the product of effective population size and generation time of the agents within each host was identical and obeyed a logistic growth function $N(t)$:

$$N(t) = \frac{N_0(1 + e^{-rT_{50}})}{1 + e^{-r(T_{50}-t)}}$$

where the timescale was in negative time and distinct for each host and $t = 0$ was the point of infection. N_0 represents the effective population size at $t = 0$, r the growth rate during the exponential growth phase of the logistic function, and T_{50} the time such that $N(T_{50})$ is half the value $\lim_{t\to-\infty} N(t)$ that $N(t)$ takes as it approaches $-\infty$. I conditioned the simulation on all lineages coalescing before $t = 0$. The complete set of such phylogenies was joined up to produce a single phylogeny for the entire simulated epidemic.

This full phylogeny was then used to generate simulated sequences using the program πBUSS [14]. Sequences consisted of 14,000 base pairs (roughly equivalent to a full influenza A genome). A strict molecular clock model with no rate variation between sites and equal nucleotide frequencies was used. Two sets of sequences were generated. The first used an unrealistically fast molecular clock with a rate of 5×10^{-4} substitutions per site per day (0.183 per site per year) while the second had a rate of 1×10^{-5} per site per day (3.65×10^{-3} per site per year). Both used the HKY substitution model [61] with a κ value of 2.718. Table 6.2 gives the parameter values actually used in the simulations.

Sequence datasets from a total of 25 simulation replicates were used for analysis. I used the within-host coalescent (WHC) method outlined in the previous chapter and sections 6.2.2-6.2.3, implemented in BEAST, to reconstruct the full phylogeny and transmission tree for each replicate, and estimate the parameters of the model that generated them. I also performed the same analysis using a blank alignment, sampling from the prior distribution only. Uninfected susceptibles were included in the analysis. For comparison, I also reconstructed the phylogeny only using a GMRF Bayesian skyride [102] tree prior. Table 6.2 also details the prior distributions used on all parameters. In this chapter I concentrate primarily on the between-host model, so the chosen priors on the within-host parameters were somewhat informative about their known values. A couple of points warrant further explanation.

Symbol	Meaning	Actual value	Prior distribution
α	Transmission kernel dispersion parameter	10	exp(10)
b	Unmodified infection rate	0.1/day	$\mathcal{U}(0,\infty)$
r	Within-case logistic growth rate	1.5/day	None[1]
N_0	N_e at time of infection	0.1	None[2]
T_{50}	Time before time of infection at which N_e achieves half its limit	-4 days	Gamma$(10,2)$
S	Ratio of $\lim_{t\to-\infty} N_e(t)$ at $-\infty$ to N_0	55.6	ln$\mathcal{N}(4,0.5)$
P^{lat}	Latent period	2 days	Gamma$(200,100)$
μ_{inf}	Mean of distribution of infectious periods	10 days	NormalGamma$(10,0.01,1,1)$[3]
τ_{inf}	Precision of distribution of infectious periods	1 days^{-2}	
	Molecular clock rate:		
	Fast clock datasets	5×10^{-4}/site/day	exp(0.1)
	Slow clock datasets and prior analysis	1×10^{-5}/site/day	None[4]
κ	HKY model transition/transversion ratio	2.718	ln$\mathcal{N}(1,0.64)$

[1] The prior probability of r is implicitly specified by the priors on T_{50} and S

[2] N_0 was fixed to its correct value of 0.1 in the analysis

[3] The slow clock analysis was also repeated with NormalGamma$(10,10,1,1)$ instead

[4] In the analyses of the slow clock datasets and the analyses sampling from the prior distribution only, R was fixed to its correct value of 1×10^{-5}/site/day

Table 6.2: Explanation of the mathematical symbols used in the simulation model, and prior distributions for their values used in analysis of the simulated datasets. Mathematical symbols are given where they appear in the text.

Firstly, in the reconstruction I assumed that all infectious periods were drawn from an unknown normal distribution with mean μ_{inf} and precision τ_{inf} and placed a conjugate NormalGamma$(\mu_0, \kappa_0, \alpha_0, k_0)$ hyperprior on μ_{inf} and τ_{inf}. The meaning of this is that τ_{inf} is gamma distributed with shape α_0 and rate k_0, and for a known value of τ_{inf}, μ_{inf} is normally distributed with mean μ_0 and precision $\kappa_0\tau_{\text{inf}}$. Initial analyses of both datasets had $\mu_0 = 10, \kappa_0 = 0.01, \alpha_0 = 1, k_0 = 1$. While this value of μ_0 is equal to the actual mean of the distribution used to generate the simulations, the low κ_0 actually means that this hyperprior is only very weakly informative about μ_{inf}. As it proved that for datasets generated with the slower clock this resulted in a systematic underestimation of the length of infectious periods (see Results), the analysis was repeated with $\kappa_0 = 10$, a modification which makes the hyperprior much more informative about μ_{inf}.

Secondly, the calculation of coalescent probabilities outlined above conditions on the time of coalescence of all lineages within each host being before the time of infection of the host (in backwards time), in common with the assumption that transmission is a complete bottleneck. I found that the estimation procedure for the parameters of the coalescent process was utterly inaccurate unless this bottleneck assumption was implicit in the priors placed on them. The reason is that the probability expressions (6.5) and (6.6) can be increased by decreasing the value of their denominators, and the denominators are very small when truncating the distribution of coalescence times actually removes a huge proportion of the probability space. The situation is that the MCMC algorithm will prefer coalescent parameters implying that coalescence while the host was infected is extremely unlikely, but this model nevertheless insists that this happened. This can be counteracted by ensuring that these parameter values have a low prior probability. The nature of the mathematics of the coalescent process used here is such that values that make the bottleneck complete cannot actually be picked, so I instead ensured that it was at least not unreasonably wide. The ratio S of the final

asymptotic value of $N(t)$ to N_0, its value at the point of infection, in the logistic model is:

$$\begin{aligned} S &= \frac{\lim_{t\to-\infty} N(t)}{N_0} \\ &= 1 + e^{-rT_{50}} \end{aligned}$$

The concerning situation is where S is small. If T_{50} is positive then S cannot be greater than 2, so I assumed it to be negative. I then placed a lognormal prior on S. This prior, combined with one on T_{50}, specifies the prior probability of r so I gave the latter no explicit distribution. I also fixed N_0 to its correct value in all simulations.

In analysing the sequence datasets generated by the slower molecular clock, the amount of genetic variation accumulated over the timescale of each epidemic was found to insufficient, for some simulations, to provide good estimates of the clock rate. As a result, this parameter was also fixed to its correct value. The same was done for the prior analysis. All MCMC chains were run for sufficiently long to give effective sample sizes of at least 200 for all numerical model parameters.

Accuracy of the reconstructed phylogenetic tree topology was assessed by counting, for each tree in the posterior sample, the number of subtree prune and regraft (SPR) moves required to take it to the correct phylogeny and taking the posterior median value of this count; I used the program rSPR [164] to determine this.

I used two methods to assess procedures by which transmission tree might be reconstructed in practice. Firstly, the posterior set of trees was summarised in a single maximum parent credibility (MPC) transmission tree, analogous to the maximum clade credibility (MCC) tree for phylogenies. The posterior distribution of parents for each case in the epidemic was calculated for each case in turn, and the parent credibility of each tree in the sample was calculated as the product of

the posterior probabilities of each link in the chain. The MPC tree is the tree in the sample that maximises this product. This was compared to the correct transmission tree, and the proportion of parents that were correctly identified calculated.

As an alternative approach I identified, for each host, the infector with the highest posterior probability, regardless of whether the result of doing this for every host actually constituted a proper transmission tree that was connected with no cycles. I calculated the proportion of parents that would be correctly identified by doing this, firstly if the actual value of the posterior probability was not considered, and subsequently for different values of a threshold probability below which inference of parental relationships would not be made.

6.2.6 Analysis of sequences from the 2003 H7N7 avian influenza outbreak in the Netherlands

Epidemiological data from the Dutch epidemic consisted of cull dates for all 241 farms, and the matrix of spatial distances between them, rounded to the nearest kilometer. (I did not have access to their precise spatial locations.) The GISAID database [16] contains sequences for isolates taken from 229 of the farms (95.0%); this consists of the HA, NA and PB2 segments in 226 cases, the HA and PB2 in 2, and the HA and NA in 1. The dates upon which these samples were taken were also available. In the absence of any other information, I assumed that a single examination of each farm took place at this time.

The HA, NA and PB2 sequences were each aligned using the MUSCLE algorithm [42]; segments which were missing were given noninformative sequences consisting entirely of the nucleotide code "N". This included entirely noninformative sequences for the twelve farms for which I had no genetic data at all; the examination date

of these was set to the cull date of the farm, a time at which it was certainly possible to acquire a sequence. The three segments were then concatenated to produce a single alignment. The 143rd codon position of the HA segment, which has been observed to cause discrepancies between reconstructed phylogenies of each segment probably as the result of convergent evolution [10], was removed. As I lacked data on the location of uninfected farms in the country, I did not include uninfected premises in the analysis, and as a result I was estimating b' (see Bayesian Decomposition), not b.

The parameters of this analysis, and the prior distributions used for them, are summarised in table 6.3. I used the SRD06 nucleotide substitution model [132] and an uncorrelated lognormal relaxed molecular clock [37]; the mean clock rate was not fixed *a priori*. The type of spatial transmission kernel function used here was the same as that used by Boender et al. [15] in their analysis of the same epidemic, determined by a logistic expression:

$$F(a_i, a_j) = \frac{1}{1 + \left(\frac{d(a_i,a_j)}{\alpha_2}\right)^{\alpha_1}}$$

where $d(a_i, a_j)$ is the distance between the farms a_i and a_j. As before, the latent period of the disease was assumed to be constant, and I placed a strong prior with a mean of two days on its length. I also followed Boender et al. in assuming that the distribution of farm infectious periods prior to the discovery of the epidemic and the implementation of control measures was distinct from that afterwards, and grouped the set of farms into "high-risk" and "low-risk" categories accordingly. The first five detected cases (F1-F5) were in the high-risk category. The hyperpriors on the distribution of infectious periods in both categories were informed by estimates from Boender et al. and Stegeman et al. [140].

I chose to regard the agent population as being made up of infected birds. As in the simulations, I assumed that the product of the effective size and the generation

time of this population within each farm underwent logistic growth, and that the same growth function was shared by all farms. Also as in the simulations, I did not estimate N_0 and instead assumed that the effective population size at the point of infection was 1, and that the generation time (the serial interval of the infection) was 3.37 days, a number derived from White and Pagano [165].

Multiple MCMC runs were performed, and the results combined using the LogCombiner utility in order to achieve ESS values over 200 for the posterior and prior probabilities, the likelihood, and all parameters listed in table 6.3.

Parameter	Symbol	Prior distribution
Transmission kernel dispersion parameters	α_1, α_2	$\mathcal{U}(0, \infty)$
Unmodified transmission rate[1]	b'	$\mathcal{U}(0, \infty)$
Within-farm logistic growth rate	r	Gamma$(20, 4)$
Product of effective population size and pathogen generation time at point of infection	N_0	None[2]
Time before infection time at which N_e achieves half its final asymptotic value	T_{50}	None[3]
Ratio of $\lim_{t\to-\infty} N_e(t)$ at $-\infty$ to N_0	S	ln$\mathcal{N}(4, 0.5)$
Latent period	P^{lat}	Gamma$(200, 100)$
Mean of distribution of infectious periods, high-risk period		NormalGamma$(7.3, 169.0, 1, 3.8)$
Precision of distribution of infectious periods, high-risk period		
Mean of distribution of infectious periods, low-risk period		NormalGamma$(13.8, 2.64, 1, 3.8)$
Precision of distribution of infectious periods, low-risk period		
Mean molecular clock rate (real space)		$\mathcal{U}(0, \infty)$
Standard deviation parameter of relaxed molecular clock (log space)		Exp(0.33)
Transition/transversion ratio		ln$\mathcal{N}(1, 0.64)$
Shape parameter of gamma distribution for between-site rate variation		Exp(0.5)
Nucleotide frequencies		$\mathcal{U}(0, 1)$
Relative clock rates for nucleotide positions 1+2 and 3		$\mathcal{U}(0, \infty)$

[1] This parameter is not the true unmodified transmission rate b that would be estimated in the presence of data on uninfected susceptibles; see the text for details.

[2] N_0 was fixed to 3.37 in the analysis

[3] The prior probability of T_{50} is implicitly specified by the priors on r and S

Table 6.3: Parameters used in the H7N7 analysis, and prior distributions on their values

6.3 Results

6.3.1 Simulations

Fig. 6.1 summarises the accuracy of the reconstruction of the transmission tree and the estimates of infection times for each host. For the latter, I saw low bias and error when the molecular clock rate was fast. However, the use of realistic sequences led to a systematic tendency to underestimate times from infectiousness to removal when the mean parameter of the probability distribution from which infectious periods are drawn was not given a strongly informative prior. It is clear from the results of the prior analysis that the effective prior distribution favours short infected periods. Re-running the analysis with an informative prior on μ_{inf} (by setting $\kappa_0 = 10$, see section 6.2.5) greatly reduced this effect, but did not entirely eliminate it.

The transmission tree was very well reconstructed when the clock was fast, with the posterior median proportion of parents being correctly identified, across the 25 simulations, having a median of 0.92 (range 0.82-0.98). For the slower clock this was considerably reduced, with a median of 0.64 (0.46-0.78). Increasing κ_0 had no noticeable effect on this (median 0.64, range 0.43-0.78). As expected, reconstruction of the transmission tree when MCMC samples were taken from the prior distribution only was extremely poor (median 0.13, range 0.08-0.18). The MPC transmission tree's median proportion of correctly identified parents was 0.96 (0.83-1.00) for the fast clock dataset, 0.71 (0.60-0.90) for initial slow clock dataset, 0.71 (0.54-0.86) for the slow clock dataset with $\kappa_0 = 10$, and 0.17 (0.1-0.28)

Table 6.4 summarises the accuracy of the procedure of picking the infector with the highest posterior probability, for no probability threshold and thresholds of 0.5, 0.8, and 0.9. It can be seen that for a threshold of 0.8, inferences are highly

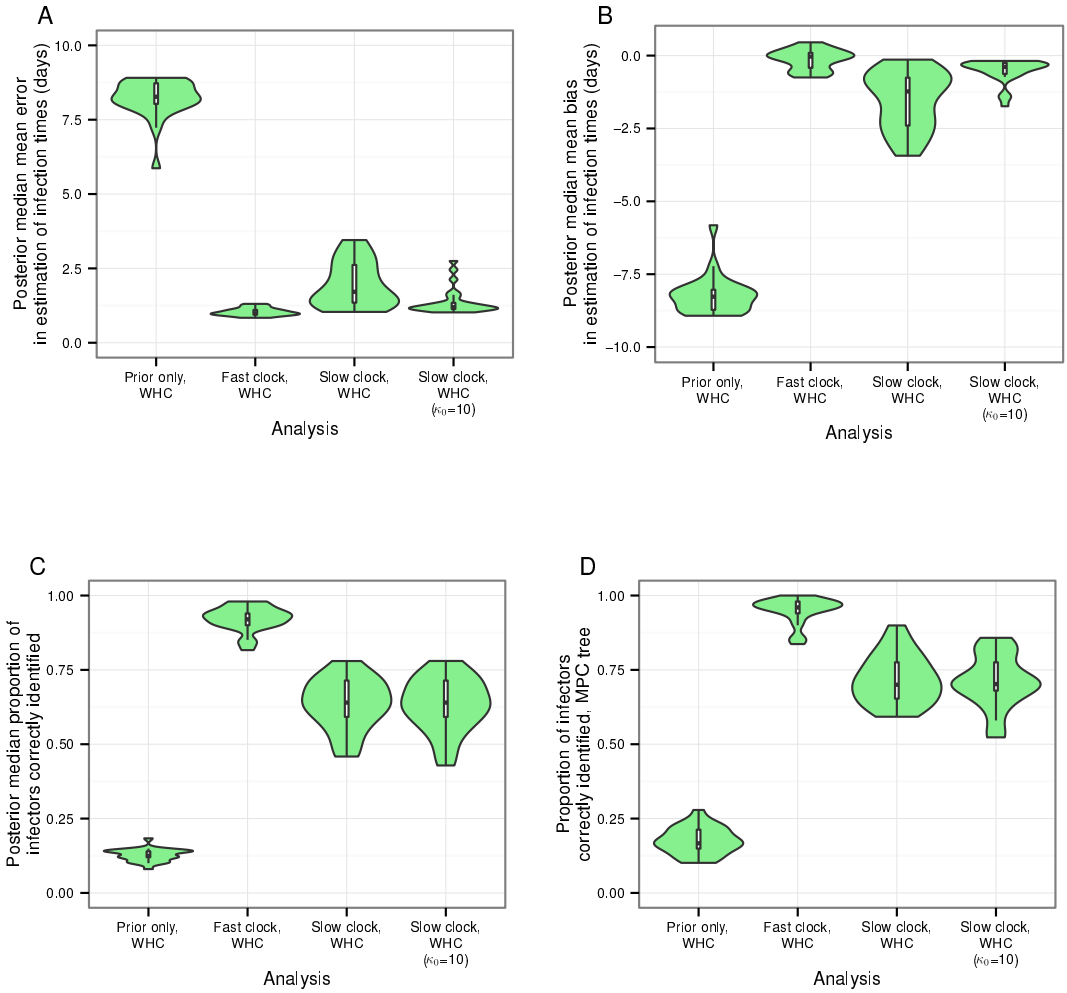


Figure 6.1: Accuracy of the reconstruction of the transmission tree. Each violin plot represents the density of a statistic over the 25 simulations. (A) posterior median of mean bias in estimation of infection dates. (B) posterior median of mean error in estimation of infection dates. (C) Posterior median proportion of hosts whose infector is correctly identified. (D) Proportion of hosts whose infector is correctly identified in the maximum parent credibility (MPC) transmission tree.

accurate even for the slow clock dataset and that the use of a value of this size leaves up to two-thirds of hosts with an inferred infector.

Analysis	Statistic	Threshold			
		None	0.5	0.8	0.9
Prior	% Parents correct	20(10, 28.5)	50(14.3, 100)	83.3(0, 100)	100(0, 100)
	% Parents inferred	100	14.8(6.0, 39.6)	6.0(2.0, 16.6)	6.0(2.0, 14.6)
Fast clock	% Parents correct	96.0(83.7, 100)	96.0(83.7, 100)	97.8(88.1, 100)	100(92.9, 100)
	% Parents inferred	100	100(95.7, 100)	90.0(76.0, 100)	84.0(68.0, 100)
Slow clock	% Parents correct	72.0(59.2, 88.0)	81.0(73.5, 89.6)	94.1(86.2, 100)	96.0(80.0, 100)
	% Parents inferred	100	81.6(54.0, 96.0)	51.0(16.3, 64.0)	38.3(12.5, 58.0)
Slow clock	% Parents correct	72.0(53.1, 88.0)	83.3(71.8, 91.4)	93.8(86.7, 100)	100(86.4, 100)
$\kappa_0 = 10$	% Parents inferred	100	81.6(54.0, 98.0)	48.9(14.3, 64.0)	38.8(8.16, 54.0)

Table 6.4: Percentage of cases with parents correctly identified by picking the infector case with the highest posterior probability for different thresholds, and percentage of cases whose parents are inferred in this way for each threshold. Numbers are median and range across the 25 simulations.

For the fast clock sequences, the phylogeny was sufficiently well resolved by the genetic data that the skyride and WHC methods performed similarly in reconstructing it, but WHC performed better when the molecular clock rate was more realistic (fig. 6.2). Error and bias in the estimates of the TMRCA of each pair of sequences was notably reduced for WHC. Using an informative prior on μ_{inf} appears to have made estimates slightly better still, although the improvement is small. The reconstruction of the topological structure of the phylogeny was also improved, with the number of SPR moves needed to take a sampled tree from the MCMC chain to the true phylogeny being consistently smaller for WHC, where the median (across the 25 simulations) posterior median number of required SPR moves was 14 (range 8-21), compared to the skyride analysis, where it was was 18 (range 11-24). The informative prior on μ_{inf} made no noticeable difference in this case.

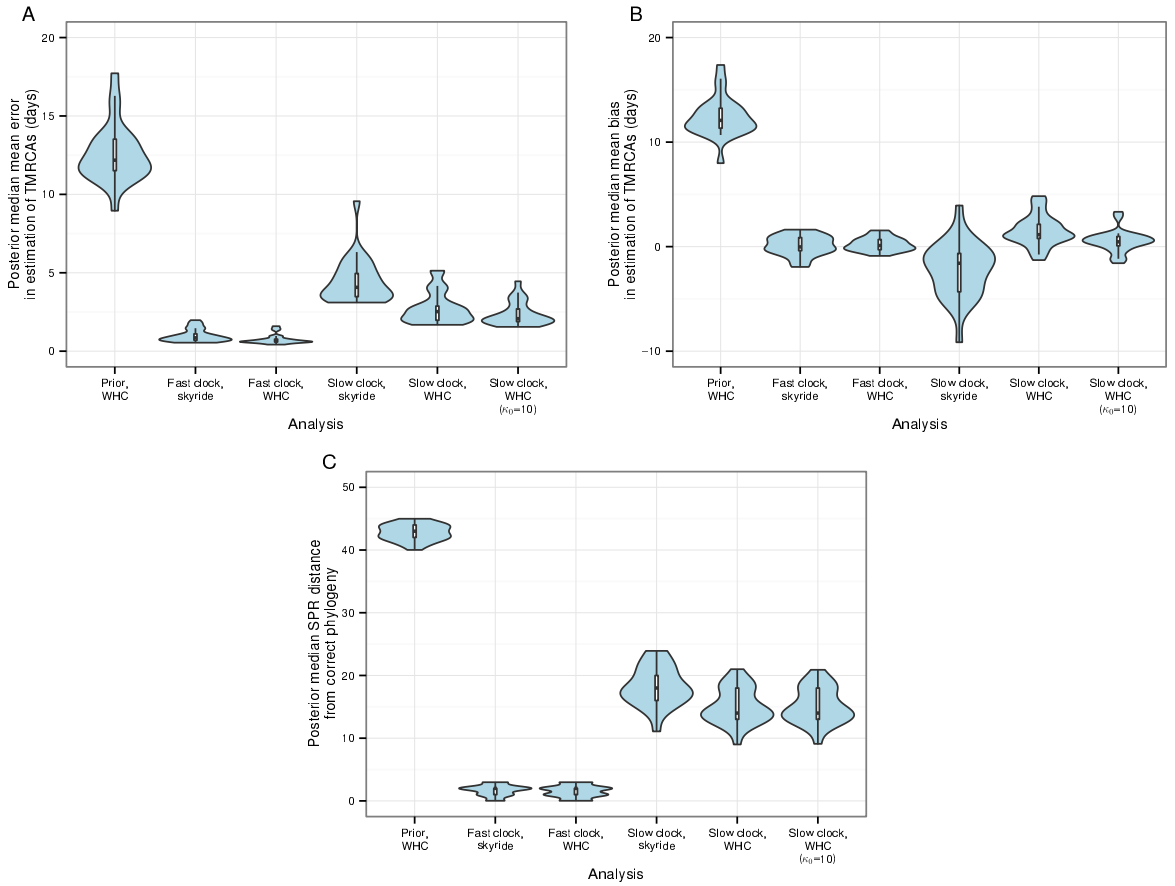


Figure 6.2: Accuracy of the reconstruction of the phylogeny. Each violin plot represents the density of a statistic over the 25 simulations. (A) posterior median of mean bias in estimation of all pairwise TMRCAs. (B) posterior median of mean error in estimation of all pairwise TMRCAs. (C) Posterior median SPR distance from the true phylogeny.

Table 6.5 summarises the posterior parameter estimates and their accuracy. Figures given are the medians across the 25 simulations. The tendency of WHC to substantially underestimate infectious periods unless an informative prior is used on the mean of their distribution is also clear here; latent periods were also slightly underestimated although the true values were always well within the 95% highest posterior density (HPD) interval. It is also noticeable that the parameters r and T_{50} of the logistic growth function describing within-host effective population size are not well estimated for the slow clock dataset, with very wide HPD intervals and also a great bias towards underestimating the value of the latter, to the extent that the 95% HPD was frequently inaccurate. On the other hand, the ratio S was recovered with much more precision and much less error and bias. These within-host parameters were in fact rather better estimated when sampling was from the prior only, but this is presumably because the prior distributions on them were chosen with knowledge of their true values.

Symbol	Meaning	Dataset	Tree prior	True value	Median	Error	Bias	95% HPD width	HPD accuracy
	Molecular clock rate[1]	Fast	Skyride	5×10^{-4}	5.11×10^{-4}	2.68×10^{-2}	2.13×10^{-2}	0.13	24
		Fast	WHC	5×10^{-4}	5.07×10^{-4}	2.19×10^{-2}	1.34×10^{-2}	0.11	23
κ	Transition/transversion ratio	Prior	WHC	2.72	2.59	4.58×10^{-2}	-4.58×10^{-2}	7.49	25
		Fast	Skyride	2.72	2.72	2.12×10^{-2}	-1.09×10^{-3}	0.11	23
		Fast	WHC	2.72	2.72	2.17×10^{-2}	5.17×10^{-4}	0.11	23
		Slow	Skyride	2.72	2.52	0.13	-7.13×10^{-2}	0.74	24
		Slow	WHC	2.72	2.52	0.13	-7.28×10^{-2}	0.76	24
		Slow	WHC ($\kappa_0 = 10$)	2.72	2.51	0.13	-7.56×10^{-2}	0.76	24
$\overline{\mathbf{P}^{\text{inf}}}$	Mean infectious period[2]	Prior	WHC	10	1.97	0.8	-0.8	0.13	0
		Fast	WHC	10	9.97	2.05×10^{-2}	-1.32×10^{-3}	0.17	24
		Slow	WHC	10	8.78	0.11	-0.11	0.29	15
		Slow	WHC ($\kappa_0 = 10$)	10	9.7	2.72×10^{-2}	-2.72×10^{-2}	0.19	24
$\sigma(\mathbf{P}^{\text{inf}})$	Standard deviation of infectious periods[2]	Prior	WHC	1	0.93	0.23	-0.14	0.59	17
		Fast	WHC	1	1.06	8.74×10^{-2}	2.39×10^{-2}	0.78	24
		Slow	WHC	1	1.16	0.31	0.15	1.75	23
		Slow	WHC ($\kappa_0 = 10$)	1	1.11	0.25	0.2	1.86	25
P^{lat}	Latent period	Prior	WHC	2	1.71	0.14	-0.14	0.25	7
		Fast	WHC	2	1.95	2.48×10^{-2}	-2.48×10^{-2}	0.26	25
		Slow	WHC	2	1.91	4.39×10^{-2}	-4.39×10^{-2}	0.26	24
		Slow	WHC ($\kappa_0 = 10$)	2	1.87	6.67×10^{-2}	-6.67×10^{-2}	0.25	25
α	Transmission kernel dispersion parameter	Prior	WHC	7	3.1	0.56	-0.56	1.53	24
		Fast	WHC	7	7.2	9.45×10^{-2}	2.89×10^{-2}	0.63	24
		Slow	WHC	7	7.22	0.14	3.12×10^{-2}	0.77	25
		Slow	WHC ($\kappa_0 = 10$)	7	7.29	0.14	4.13×10^{-2}	0.78	25
b	Unmodified transmission rate	Prior	WHC	0.1	9.02×10^{-2}	0.13	-9.77×10^{-2}	5.76	24
		Fast	WHC	0.1	0.11	0.22	0.13	1.32	24
		Slow	WHC	0.1	0.11	0.27	0.15	1.73	24
		Slow	WHC ($\kappa_0 = 10$)	0.1	0.11	0.25	0.12	1.66	24
r	Within-host logistic growth rate	Prior	WHC	1	0.83	0.17	-0.17	3.14	25
		Fast	WHC	1	1.06	0.15	5.8×10^{-2}	0.99	24
		Slow	WHC	1	2.76	1.76	1.76	5.96	23
		Slow	WHC ($\kappa_0 = 10$)	1	2.49	1.49	1.49	5.73	25

Continued on next page

Symbol	Meaning	Dataset	Tree prior	True value	Median	Error	Bias	95% HPD width	HPD accuracy
T_{50}	Time at which within-host population size is half its final value	Prior	WHC	−4	−4.35	0.36	−0.35	7.3	25
		Fast	WHC	−4	−3.6	0.75	0.4	3.24	24
		Slow	WHC	−4	−1.38	2.62	2.62	3.47	12
		Slow	WHC ($\kappa_0 = 10$)	−4	−1.5	2.5	2.5	4.18	19
S	Ratio of final within-host population size to size at infection	Prior	WHC	55.6	36.83	0.34	−0.34	1.34	25
		Fast	WHC	55.6	44.55	0.22	−0.2	1	25
		Slow	WHC	55.6	46.61	0.16	−0.16	1.11	25
		Slow	WHC ($\kappa_0 = 10$)	55.6	43.63	0.22	−0.22	1.11	25

[1] Molecular clock rates were not estimated for runs on the slow clock dataset

[2] Infectious periods were drawn from a normal distribution with the "actual values" given here as mean and standard deviation. Error and bias were, however, calculated using the mean and standard deviation of the actual set of estimated periods in each MCMC state.

Table 6.5: Estimates of simulation parameters from the various analyses. The median of the posterior median, relative error in the median, relative bias in the median and relative 95% HPD widths over the 25 simulations are given, along with the number out of 25 simulations that the correct value was contained within the 95% HPD interval. Where estimates are not given for a particular analysis, this parameter was either fixed to its correct value or, in the case of WHC-related parameters in skyride analyses, not part of the analysis. Mathematical symbols are given where they are referred to in the text.

6.3.2 Analysis of sequences from the 2003 H7N7 avian influenza outbreak in the Netherlands

The MPC transmission tree can be seen in fig. 6.3. It can be seen that most of the inferred transmissions have a quite low posterior probability. In fact, if a posterior probability threshold of 0.5 was used to infer transmissions, conclusions would be drawn about only 83 farms (34.4%), with this dropping to 23 (9.5%) for a threshold of 0.8, and only 7 (2.9%) for a threshold of 0.9. None of the five "high-risk" farms met the 0.5 threshold, and the posterior probability that the index case was among these five was only 0.36. This is the reason why in the MPC tree the presumed index farm F1 is not correctly identified, and while it and the other five high-risk period farms are close together at the start in the transmission chain, they are interrupted by one farm from the low-risk period. This is farm F6, for which a virus was sampled on the very day that control measures were implemented. The posterior median date of the first infection was the 15th February, 2003, almost two weeks prior to detection, with the 95% HPD ranging from the 7th until the 18th. This is highly consistent with the modelling work of Bos et al. [17], which had no genetic component. The orange-bordered nodes in the tree are the twelve farms for which no sequence is available. Notably, the procedure placed them amongst their geographical neighbours. Figure 6.4 is the MCC phylogeny, with branches coloured by individual farm. It should be noted that the branch colourings in this figure do not reflect a history of the epidemic that is particularly representative of the posterior sample of transmission trees; they are simply the colourings of the phylogeny from the posterior with the highest clade credibility.

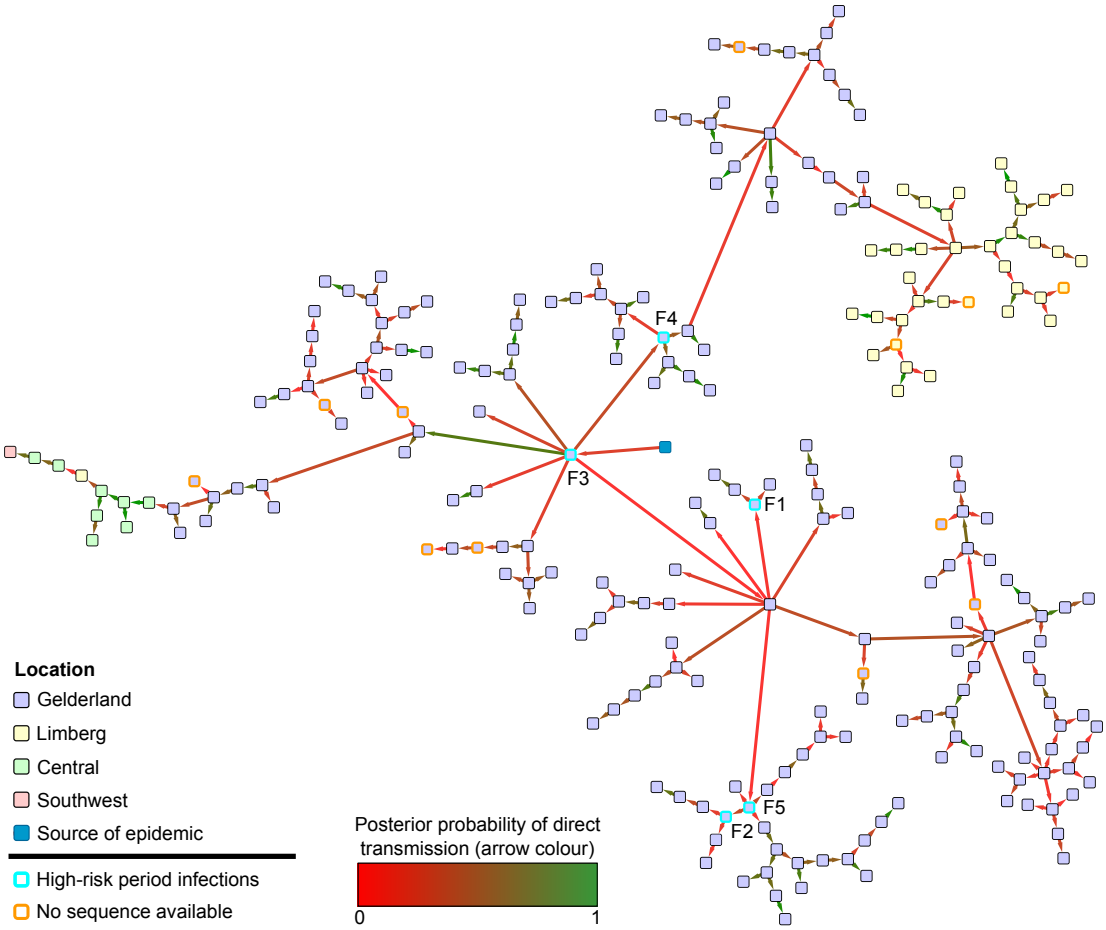


Figure 6.3: Maximum parent credibility transmission tree for the H7N7 outbreak. Nodes represent farms and are coloured by geographical region. Arrows represent direct transmissions and are coloured by the posterior probability of this particular direct infection. The cyan-bordered nodes, which are also labelled with farm ID numbers from previous literature [10], are were detected during the "high-risk" period before the implementation of control measures. Orange-bordered nodes are farms for which no sequence was available.

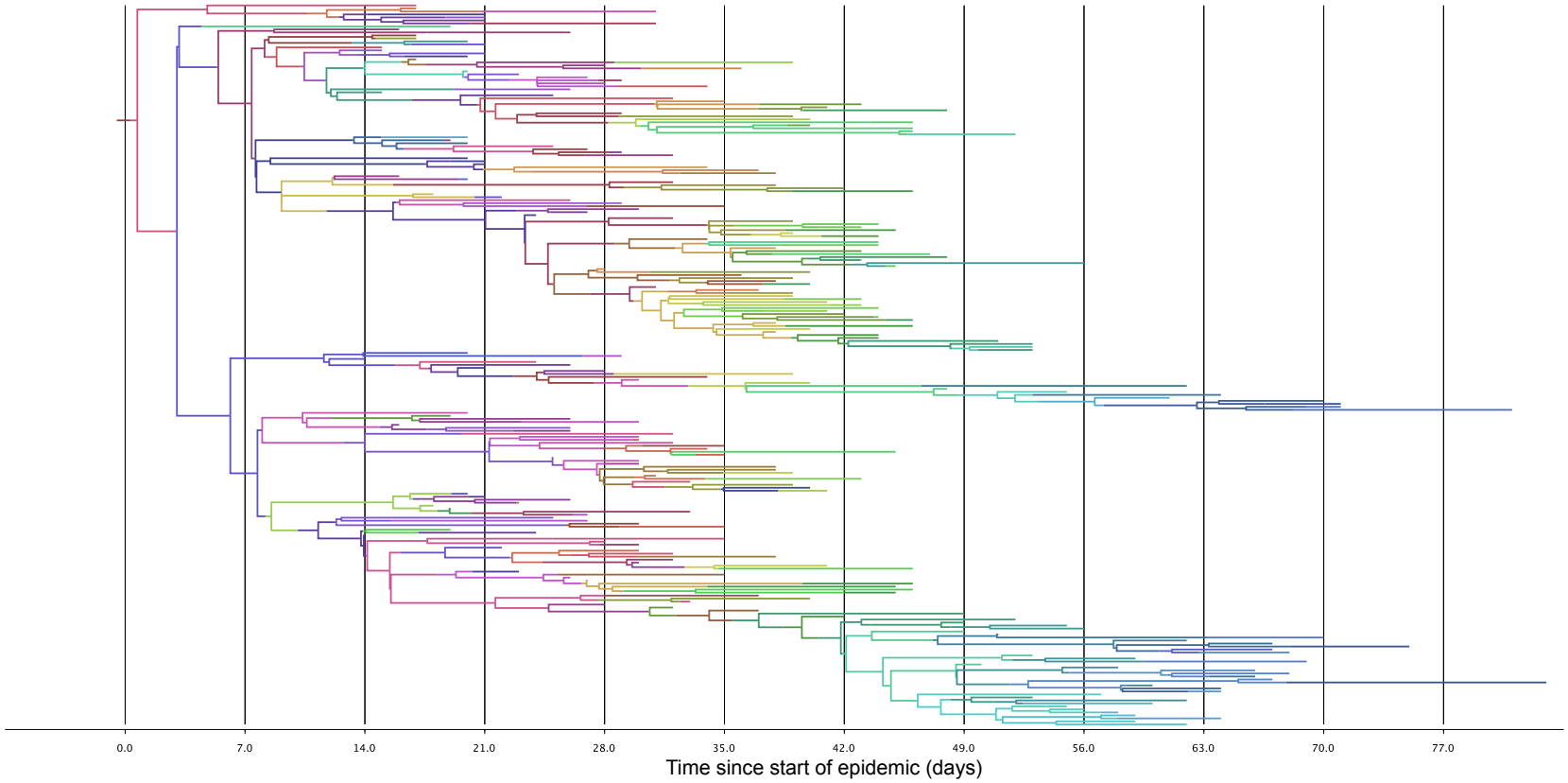


Figure 6.4: Maximum clade credibility phylogeny for the H7N7 outbreak. This is the actual sampled tree with highest clade credibility from the posterior set; branch lengths have not been adjusted. Branches are coloured by farm; colour changes along branches reflect infection events.

Table 6.6 summarises the parameter estimates. Of note, while I used an extremely informative prior distribution with a mean of two days on the length of the latent period, the estimate was still considerably shorter (posterior median 1.47 days, 95% HPD 1.26-1.69). The estimated mean infectious period in the low-risk period did not deviate greatly from the prior expected value of 7.3 (posterior median 7.42 days, 95% HPD 6.93-7.93). On the other hand, the posterior distribution for the mean for the high-risk period had a median of 8.11 days (95% HPD 5.19-11.9), considerably smaller than the prior expected value of 13.8. Compared to the prior expected value of the precision of the distribution in both high-risk and low-risk periods of 0.263 days^{-2}, the estimated precisions were lower, with posterior medians of 0.131 (1.61×10^{-2}-2.37) for the high-risk period and 8.13×10^{-2} (6.26×10^{-2}-0.104) for the low-risk period. The parameters of the within host population function suggested that the effective size of the infected population rose very quickly towards its asymptotic value, achieving values extremely close to it within a day or so. If the median estimates were used, this asymptotic population size was 11.4 times the value at the point of infection. While this behaviour would not seem to reflect the likely course of an epidemic within a flock, it should be remembered that the within-farm model was extremely simplistic in this example. The estimated parameters of the transmission kernel are consistent with the maximum likelihood figures from Boender et al. [15] (work which also had no genetic component), which were, in my notation, $\alpha_1 = 2.1$ and $\alpha_2 = 1.9$. As the analysis in that paper did have access to data on every farm in the Netherlands, they were able to estimate b (in my notation) to be 2×10^{-3} days^{-1}. My estimate of b' is larger, as expected, but they are not directly comparable. For further consideration, see section 6.4.

Parameter	Symbol	Median value (95% HPD)
Transmission kernel parameters	α_1	2.15 (1.70, 2.72)
	α_2	2.00 (0.92, 3.39)
Unmodified transmission rate	b'	2.78×10^{-2} /day (1.18×10^{-2}, 5.33×10^{-2})
Within-farm population growth rate	r	7.04 /day (4.64, 9.96)
Time before infection time at which N_e achieves half its final asymptotic value	T_{50}	-0.33 days (-0.45, -0.23)
Latent period	P^{lat}	1.47 days (1.26, 1.69)
Mean infectious period (high-risk period)		8.11 days (5.20, 11.85)
Standard deviation of infectious periods (high-risk period)		2.76 days (0.65, 7.88)
Mean infectious period (low-risk period)		7.42 days (6.93, 7.93)
Standard deviation of infectious periods (low-risk period)		3.51 days (3.10, 4.00)
Mean molecular clock rate		2.68×10^{-5} subs/site/day (2.26×10^{-5}, 3.11×10^{-5})
Standard deviation of molecular clock rates		1.34×10^{-5} subs/site/day (2.24×10^{-6}, 2.40×10^{-5})
Transition/transversion ratio		**Positions 1+2:** 7.09 (4.73, 9.98)
		Position 3: 9.01 (5.58, 13.3)
Shape parameter of gamma distribution for between-site rate variation		**Positions 1+2:** 3.77×10^{-2} (1.00×10^{-3}, 9.43×10^{-2})
		Position 3: 0.230 (1.02×10^{-3}, 0.661)
Nucleotide frequencies		**A:** 0.334 (0.321, 0.345)
		C: 0.188 (0.178, 0.198)
		G: 0.249 (0.237, 0.259)
		U: 0.230 (0.219, 0.240)
Relative clock rate parameter		**Positions 1+2:** 0.854 (0.753, 0.936)
		Position 3: 1.29 (1.13, 1.49)

Table 6.6: Estimates of parameters from the H7N7 outbreak, posterior median and 95% HPD interval

6.4 Discussion

I provide here a novel method for simultaneous reconstruction of both phylogenies and transmission trees, fully incorporated into the existing BEAST package. Being part of an established package has the advantage that users of the method have access to existing models and methods for, for example, relaxed molecular clocks, ancestral sequence reconstruction, coalescent population models, and marginal likelihood estimation, without the need for extra programming work. The prior probability decomposition outlined above is also very flexible, allowing for many different distributions of infectious periods and models of spread between hosts. As such, this represents a substantial evolution and generalisation of the earlier work of Ypma et al. [173], who gave two separate and considerably different decompositions designed specifically for two particular datasets, and also were not able to accommodate an assumption of homogeneity of infectious periods.

The method works by annotating the internal nodes of the phylogeny here in the same way as Didelot et al. [35], the "colours" of that paper correspond to my partition elements. As mentioned in chapter 5, I do not, however, assume as they do that each element contains only a single tip; this procedure can deal with multiple sequences for samples taken from the same host, under the assumption that the host was infected only once. (If this may not be true, it might be more appropriate to treat the two introductions as separate "hosts", particularly in an agricultural scenario, and if recombination or reassortment between two infecting strains can be ruled out.) This would be of use in, for example, the study of HIV, where multiple samples are often taken from the same patient over the course of treatment [159]. I showed in the previous chapter that it is impossible for an MCMC procedure to fully explore the space of transmission trees without letting the phylogeny vary if the latter has more than two tips. I also established that varying the phylogeny does indeed allow the algorithm complete access to this

space. This is important, as the short timespan of phylogenies from epidemics and outbreaks often results in phylogenies that are not particularly well resolved, and as a result, a two-step procedure may make it impossible to infer many plausible transmission histories. While in some circumstances access to the full space of transmission trees may not be necessary because some are very implausible (it is unlikely, for example, that the last case in an outbreak lasting months was infected by the first), which trees are implausible will vary greatly depending on the nature of the pathogen. It would be reasonable to rule out direct transmission between two individuals if their infection dates were separated by years if the pathogen was influenza, but not if it was HIV. Therefore, I considered it important, in designing a method intended to be general and flexible, to allow access to every single transmission tree. A simultaneous procedure such as this is to be preferred to the option of running a separate fixed-tree analysis on each of a set of trees from a Bayesian posterior sample for reasons of computational time.

I found here that the WHC method was superior in estimating both the topology structure of the phylogeny and its node heights than an analysis using the skyride tree prior; the latter, which assumes all lineages belong to a single, freely-mixing population, is amongst the most frequently employed current methods for the reconstruction of time-resolved phylogenies. This adds weight to the concerns I expressed in the introduction to chapter 5 regarding two-step methods that use a phylogeny or set of phylogenies estimated by another method as input for epidemiological reconstruction; the assumptions under which those trees were made may violate the population model that the epidemiological inference is using, and as a result they may not be accurate. As I have here developed a more accurate tree prior for an epidemic situation, I would recommend that the WHC be used for reconstruction of the phylogeny of suitable datasets even if the transmission tree is not of interest.

A frequent concern surrounding analyses of this sort has been the question of

unsampled cases. Some progress has been made in dealing with this issue recently in the non-phylogenetic methods [75, 104], and in a recent paper, Numminen et al. [107] outlined a novel two-step, importance sampling method for the investigation of transmission trees using potentially sparsely sampled data based on a fixed, maximum-parsimony phylogeny. I go some way to addressing this problem in a one-step process because, as demonstrated, this method can include epidemiological information for known clinical cases for which no sequence is available. Scenarios of this sort, indeed, provide another reason to prefer a one-step approach; as a standard phylogenetic analysis is unaware of any epidemiological information other than dates of sampling, it has no information to use in placing a noninformative sequence in the tree. The position of a corresponding tip in a fixed phylogeny used as input for a two-step method will be effectively random. This method can, instead, use epidemiological data such as the location of the case, as well as a prior or hyperprior on the time from infection to noninfectiousness, to place these with more certainty. It can be seen from the reconstructed transmission tree (fig. 6.3) from the H7N7 epidemic that the farms for which no sequence was available are placed amongst their geographical neighbours, which would be expected unless there was a particular reason to believe otherwise.

This is obviously not a complete solution to the problem; more challenging is the issue of the identification of unknown unsampled hosts in the transmission chain, and the quantification of the number of them. This is the principal limitation of this method, and further work is needed to address it. Two approaches have been suggested previously [35, 104], both of which could be accommodated as a modification to the WHC. The first is to create a pool of unsampled cases, of variable size, and use reversible-jump MCMC (rjMCMC) [55] to add and subtract from it. Internal nodes in the phylogeny can then be assigned to elements of this set, obeying the rules about connectedness but disregarding that about each partition element containing a tip. The second is to allow hosts to "indirectly"

infect others even after they ceased to be infectious. The assignment of a node to a particular partition element would no longer indicate that the lineage represented by that node was actually present in the host represented by the tip in the same element, but just that it was infected by that host before it entered any other sampled host. I suggest a third option, which is to allow the assignment of internal nodes to no host at all. The mathematical framework would require modification; for example, an expression would be needed for the probability of the infection of a host from an unknown source.

The assumption that transmission is a complete bottleneck is hard to relax, as one of the fundamental principles of the correspondence between transmission trees and partitions, that the nodes in a partition element form a connected subtree of the whole phylogeny, must then be discarded as the common ancestral node of two nodes in the same host may be outside that host. The realism of this assumption is often unclear and will vary from pathogen to pathogen; while the bottleneck has been found to be quite loose for individual-to-individual transmission of influenza [70], this may be less true when, as in this example, transmission is between farms [10]. For other organisms, such as HIV [80] and hepatitis C virus [21], it has been found that the number of transmitted variants is usually very small between individuals.

Treating the infection status of a host upon examination as part of the data, rather than as background information, greatly simplifies the mathematics surrounding infectious periods. It has other consequences. On the positive side, it opens up the possibility of including the results of genuinely negative examinations as data, in a way that does not involve adjusting individual prior probabilities for infection times. Such a negative examination must be as near as possible to conclusive, however; the absence of clinical symptoms is certainly not sufficient. On the negative side, it means that, outside the situation where no infections persist after a sequence is acquired (a special case which I used in my simulations),

an algorithm to sample from the *prior* probability distribution must be able to vary the number of tips in the tree, a very non-standard procedure which is not implemented currently in BEAST or any other commonly-used package.

As outlined in Methods, the parameter b, the underlying rate at which a susceptible host becomes infected by a single infectious host before modification by the function F (the transmission kernel in these examples), cannot be estimated unless the set of uninfected susceptibles can be described. This may not be possible in many cases; in my H7N7 analysis I did not have such data and hence could only estimate b', the value that b would take if no susceptibles remained at the end of the epidemic. If it is not feasible to acquire this information and this parameter is still of particular interest, then a method to estimate the contribution of the set of uninfecteds is needed.

The WHC method was very successful in recapturing the epidemiological parameters of the simulations where sequences were generated by a fast clock, in which case the level of genetic diversity was such that there was little uncertainty in the phylogeny. Moving to a more realistic level of diversity decreased the accuracy of estimates considerably, and this illustrates the importance of using existing information, be it genetic or epidemiological, when configuring an analysis of this type. In particular, the results of the simulation analysis showed a clear bias towards underestimating the infectious periods of hosts. The reason for this is that the kind of within-host phylogenies that maximise the probability expressions (6.5) and (6.6) with an increasing within-host population are those that have only short periods in which the tree has more than one lineage, and where those short periods are close to the time of infection. The probability each such phylogeny is therefore increased, all else being equal, by moving the infection time towards the tips. The priors placed on the lengths of infected periods and on the parameters of the within-host coalescent process are not, in this mathematical framework, consistent, with the result that the effective priors are not the distributions chosen by the

user. This a similar phenomenon to the effect of placing a "calibration density" on the root height of a coalescent tree that conflicts with the prior distributions for coalescent parameters [63], although the situation is considerably more complex as the WHC deals with multiple coalescent trees with mobile tip dates. Further analytical work may be able to provide consistent prior distributions. This effect in any case is overwhelmed by sufficient genetic data (as in the fast clock dataset) and can be largely mitigated by placing a suitably informative prior on the length of infectious periods. A clear implication of this is that genetic data should not be relied upon to estimate infectious periods on their own if other information is available to inform such a prior. (I note that this bias does not affect estimation of who infected who, as the accuracy of transmission tree reconstruction did not significantly change when I changed the prior on μ_{inf}.) In a similar vein, the lack of genetic diversity in the slow clock dataset meant that in some cases the molecular clock rate itself could not be reliably estimated and had to be fixed to its known value. In an actual epidemic situation it would seem perfectly reasonable to do this, using a rate derived from older data, unless it was clear that the pathogen in question was novel.

The concerns about the use of uninformative priors on infectious periods led me to use informative distributions taken from previous literature in the H7N7 analysis, and I continued the practice from the simulations of using highly informative priors on latent periods as preliminary work had shown that these tended not to be well estimated using uniform priors. Some analysis results nevertheless deviated from what would be expected under the prior distributions; in particular the estimated mean infectious period during the high-risk period, and the estimated latent period, were both considerably smaller than their prior expected values. While it is possible that these underestimates are at least partly the result of the bias that was noted in the simulations, the difference is considerably more extreme than anything observed there. While this analysis agrees that the infectious period

in the index case may have been around two weeks, the genetic data seems to suggest, contrary to previous work [15, 140], this was not the case for the remaining high-risk farms. In the case of the latent period, the MCMC never actually sampled a value of two days or more at all, which previous work had assumed *a priori* as its length. While it is true that the assumption of a single latent period for all farms is a considerable simplification, this still suggests that the phylogeny is simply unable to accommodate a situation where all latent periods are of two days or greater. This analysis also suggested much greater variation in the lengths of infectious periods than had been previously estimated, in the low-risk period at least. These three observations suggest possible insights that genetic data can provide have not been apparent in traditional analyses.

The other model parameters that are not well recovered in the slow clock simulation dataset are those of the within-host coalescent process. This chapter has concentrated on the between-host model, and if the WHC method it is to be used to investigate within-host dynamics in detail then further work is needed. It may be that the situation is improved if multiple sequences are taken from the same host. The estimated parameters of the logistic growth function for H7N7 should certainly not be overinterpreted, for several reasons. First, it is a gross oversimplification to assume that the infected population of each farm grew according to the same function, especially when the farms infected in this epidemic ranged from hobby farms to large agricultural facilities. Secondly, the "effective number of infections" will differ from the true size of the infected population due to the violation of assumptions made in the coalescent process. While the WHC is designed to deal with violations of the assumption of homogeneous mixing of lineages that in fact infect separate hosts in an epidemic, lineages would not be expected to mix freely within farms (or indeed host organisms) either. It has also been shown that if the population in a coalescent model is treated as being made up of infected individuals, the relationship of effective population size to prevalence

is not straightforward [48, 155]. Lastly, logistic growth may be too simplistic a model of the population. It was picked here because it is clearly a better fit to growth within a closed population such as a farm than a constant population size or exponential growth, but true dynamics are no doubt more complicated still.

Even the "slow clock" simulation dataset was intended to represent the full genome of influenza A, one of the most fast-evolving pathogens that is likely to cause an outbreak of this type. The resolution in the reconstruction of the transmission tree for the H7N7 outbreak could be increased if sequences for the remaining segments of the genome were available; consistently higher posterior probabilities for infectors were observed in the simulation analyses. As the short timescale of an epidemic already places a limit on the amount of information that can be gleaned from genetic data, I would suggest that resources be expended to sequence as much of the pathogen genome as possible in a situation of this sort.

In conclusion, what I have demonstrated in this chapter and the last is both a new phylogenetic method for the analysis of genetic data taken from outbreaks and epidemics, and a new transmission tree reconstruction method. For phylogeny reconstruction I have developed a population model that is more realistic than the assumptions of freely-mixing lineages that are made in the most widely-used current methods. For transmission tree reconstruction, I have advanced the development of models that accommodate within-host diversity with a procedure that maintains the previously-noted correspondence between transmission trees and the annotation of internal nodes in a phylogeny while exploring the full space of phylogenies, which is required to allow access to the full space of transmission trees.

Chapter 7

Summary and discussion

In this chapter, I summarise the work presented in this thesis, suggest some directions for future research of a more general nature than were discussed in the individual chapters, and give some concluding remarks.

7.1 Thesis summary

Chapter 2 is, perhaps, a "current standard" analysis. It utilised up-to-date methods and available sequences to investigate the geographical and temporal dynamics of FMDV serotype SAT 2, as well as its historical movements between hosts. There was a particular focus on the 2011 outbreaks, to which I applied a novel application of the Markov jumps transition reconstruction method to give the posterior distribution for the origin of each. For data I simply used every available SAT 2 VP1 sequence. This effectively opportunistic approach likely introduced bias to, at minimum, the skyride reconstruction, and this concern motivated the approach of the next two chapters.

Chapter 3 is a more comprehensive investigation of the biasing effects of sampling schemes on phylodynamic and phylogeographical reconstructions than has previously appeared in the literature. It considered a total of eight different demographic scenarios, and found that while sampling epidemiologically unusual pathogens disproportionately often appeared to improve skygrid reconstructions in some of them, this was not consistently true and as a result I cautioned that it may not be an advisable approach in general. On the other hand, it is always preferable to sample locations and time intervals with equal probability rather than to attempt to weight by the EPS in that time and place, and this applied both to skygrid reconstruction and discrete-traits phylogeography. I propose this as a baseline recommendation for the design of future studies. It is a convenient one, as weighting by EPS is not a straightforward matter. I also showed that there was not a great amount to be gained by increasing sample sizes above around 200 for skygrid reconstruction, but that this was not true for phylogeography, and that sampling a large number of contemporaneous sequences from a single location introduces a spurious "bottleneck" effect.

I took the insights gained in chapter 3 back to a sequence analysis of FMDV in chapter 4. The samples for the analyses in that chapter were selected with much more care than they had been in chapter 2, and I also repeated the sample selection procedure many times in order to investigate whether features of the reconstruction might have been due to stochastic sampling effects (see appendix A). I first analysed data from the complete type O serotype, demonstrating that the Cathay topotype does indeed appear to have a faster mutation rate than others and that some viruses currently extant in South America are most likely the descendant of improperly inactivated vaccines used in the latter half of the 20th century. I then moved on to two more geographically restricted analyses, of the SEA and ME-SA topotypes. I used the GLM framework to show that the former appears to be spread solely through the cattle trade, but that the picture is less clear for

the latter. The importance of Myanmar as a reservoir of the SEA topotype was also confirmed. I also found that, as would be expected, cattle appear to be the main reservoir of both topotypes.

The remainder of the thesis, chapters 5 and 6, concerns itself with the development and implementation of a new method for the reconstruction of the transmission tree of the epidemic simultaneously with the phylogeny. Chapter 5 shows how transmission trees can be seen as partitions of the node set of a phylogeny under certain rules, and then introduces a MCMC framework by which the spaces of both types of tree can be simultaneously explored by applying such a partition to the phylogeny. Chapter 6 completes the picture by giving an example of how the posterior probability of both trees given a set of sequence data can be calculated, and test the procedure on both simulated and real data. The result is a general mathematical framework within which transmission tree inference can be performed using genetic data, one that is not specific to any particular pathogen or model of transmission. It is also fully integrated in BEAST, one of the most commonly-used existing packages for time tree inference.

7.2 Future directions

7.2.1 Sequence analysis of FMDV

An obvious way to carry this thesis' work on FMDV forward would be to apply modern phylogenetic methods to those serotypes and topotypes which have not yet had this treatment. The main serotype candidates are A, Asia-1, and perhaps SAT 1; C and SAT 3 are now very rarely occurring viruses and the amount of available historical genetic data is small. The African topotypes of serotype O are

also good candidates, although there are many potential geographical gaps in the available data.

Ultimately, all retrospective analyses of FMDV are limited by the patchy nature of the historical data. While in chapter 4 I made an attempt to choose a set of sequences for analysis in a methodological way motivated by chapter 3, this could only be taken so far before gaps in historical data become evident. On a more basic level, there is simply no data available at all for some locations, such as many African countries in chapter 2. There is, of course, no going back. Researchers cannot retrospectively produce a representative sample of the global population of FMDV (or any other a pathogen of interest) for a period in the past, and studies of historical epidemiology will always be constrained by this. However, as the technological barriers to a sequence collection effort on a very large scale disappear, a well-organised surveillance infrastructure might be able to ensure that the available data for FMDV from the next twenty years are much more comprehensive and representative than those for the last twenty.

7.2.2 Sampling strategies for phylogeography and phylodynamics

The issues addressed in chapter 3 are in some ways only the tip of the iceberg, and many questions remain. An effectively infinite variety of scenarios could be simulated, and indeed thorough study design could include this as a step.

An issue that I did not have time to explore in the chapter regarded the effects on phylogeography analysis when some locations are either entirely missing or heavily oversampled. In the former case the issue is what the effect is on the inferred movements between other locations, particularly neighbours. In the latter, the question is whether rates, or reconstructed transition counts, are biased upwards

if they involve the oversampled location. The analyses in chapter 2 and chapter 5 were both missing some countries, and in chapter 5 the number of sequences included from a particular country was supported as a predictor for the ME-SA topotype. Data limitations of these kinds are quite common in phylogeographical studies

Chapter 3 used the discrete traits model of spatial diffusion. The continuous model is less widely used due to its more stringent data requirements [95], and in fact, some of its limitations due to sampling are rather obvious. For example, if all sequences are from a single area but in fact represent multiple introductions to that area, then the analysis will, wholly wrongly, still pick a root location within it. Biased sampling may, however, have other, more subtle effects on the reconstruction of diffusion processes using this method, and this warrants investigation.

The principle behind chapter 3 was to design a sampling strategy to be used with existing models that minimises sampling bias. These existing models, of course, base their inference on oversimplistic assumptions (such as random mixing, and the treatment of discrete trait values as independent of population structure) and were not designed with sampling bias in mind. The alternative is to develop more realistic models which take the nature of the sample into account. The BASTA approximate structured coalescent model recently developed by De Maio et al. [30], which links the population model and migration models, shows that researchers in this field are beginning to come to terms with the issue. The authors demonstrate that inference using the CTMC discrete model, the current standard, is biased by the sampling process but that their approach, in which the location of sampling is treated as background information rather than part of the data, does not have this problem. (I analysed the host transition data from chapter 4 using BASTA, and the results are in appendix B.)

7.2.3 Transmission tree reconstruction

Ways in which future technical approaches to transmission tree reconstruction may deal with current limitations are discussed in section 6.4. On a practical level, however, it must be acknowledged that rigorous testing of any method of this sort on outbreaks in animal populations (and indeed also in human populations, as outbreaks in which it is possible to identify a large proportion of cases are unusual) is hindered by the fact that such events are rare in locations where the resources for comprehensive sampling would be available. The most suitable real datasets are from 2001 [26, 105] and 2003 [171], long before any of these procedures began development and also before it would have been possible to rapidly acquire sequences even if they had been available. The utility of these procedures during an emergency is perhaps not completely proven; while the tools now exist to begin to analyse an outbreak as soon as it is detected, it also remains to be seen how quickly the infrastructure of an affected country would be able to provide sequences in such an event. Scope exists for a simulation study on the performance of these methods in inferring transmission links under emergency conditions when the outbreak is only partially revealed, and how short the period from detection of infection to the availability of a sequence would need to be for them to be useful. In any case, however, these tools would be available for a retrospective analysis once the emergency is over, in order to aid forensic investigation of what happened.

The lack of comprehensive genetic datasets from actual outbreaks has not hindered development of these methods, however, as many of the most recently-published papers on this subject have concentrated on endemic disease [104, 107]. This is an important development for epidemic analysis as well, because the testing of methods on real data of any sort is essential if inference is to be relied upon in an emergency situation, and because the problems involved in applying such

procedures to endemic pathogens where the infected population is not well revealed are similar to those involved in handling epidemic sampling which is less than comprehensive. This can enable transmission tree reconstruction for epidemics occurring in resource-poor settings, or in richer settings before the full extent of the event becomes clear.

7.3 Concluding remarks

The era of cheap and fast nucleotide sequencing has enormous implications for science and indeed for wider society. In infectious disease epidemiology, it has the potential to allow investigation of the temporal and spatial dynamics of infectious agents at a resolution that would previously be unimaginable. The amount of available data is already increasing at a rapid rate, and there is increasing scope for very large scale collection projects. It is not inconceivable that a day is coming when a sequence will be acquired as a matter of course from every detected clinical case of an infectious disease; such efforts already exist for HIV [92]. As that virus infects humans and almost invariably causes fatal disease if untreated, it is not surprising that it is one of the first for which resources have been expended for such an effort. As sequencing becomes cheaper and more ubiquitous, however, it is to be expected that the net will be cast wider. It is true that "every detected clinical case" is a rather more comprehensive dataset for some pathogens than others, as most cases of less serious human or animal infections are never seen by a clinician or veterinarian, but there nevertheless remains scope for a vast increase in available information in almost every case. In agriculture, commercial interests may prompt such initiatives in order to investigate how the financial burden of livestock infections can be lessened, and the economically devastating effects of outbreaks of certain pathogens may also prompt large-scale sequencing projects initiated by governments.

The incoming glut of sequences means that the field of molecular epidemiology is one that is currently in transition. Procedures must be revised to be able to answer old questions with much larger amounts of data, and to approach questions that would never have been answerable in previous decades. The entire emerging field of phylodynamics, indeed, is part of the latter category. The methodological work in this thesis has contributed to advancing the field in both of these directions. If well-established methods are to be used on a large quantity of data, the question of exactly which isolates to collect, sequence and include, which has rarely been a major concern before, must be addressed, as I did in chapter 3. Transmission tree reconstruction of the type I described in chapters 5 and 6, on the other hand, is something that is only possible in the new era. Along the way, I have also provided new insights into the global dynamics of the economically important livestock virus, FMDV.

Chapter 8

Bibliography

1. Abdul-Hamid N.F., Hussein N.M., Wadsworth J., Radford A.D., Knowles N.J. and King D.P. (2011) Phylogeography of foot-and-mouth disease virus types O and A in Malaysia and surrounding countries. *Infection, Genetics and Evolution*, **11** (2), 320–328.
2. Ahmed H.A., Salem S.a.H., Habashi A.R., Arafa A.A., Aggour M.G.A., Salem G.H., Gaber A.S., Selem O., Abdelkader S.H., Knowles N.J., Madi M., Valdazo-González B., Wadsworth J., Hutchings G.H., Mioulet V., Hammond J.M. and King D.P. (2012) Emergence of foot-and-mouth disease virus SAT 2 in Egypt during 2012. *Transboundary and Emerging Diseases*, **59** (6), 476–481.
3. Aldrin M., Lyngstad T.M., Kristoffersen A.B., Storvik B., Borgan Ø. and Jansen P.A. (2011) Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *Journal of The Royal Society Interface*, **8** (62), 1346–1356.
4. Ayebazibwe C., Mwiine F.N., Tjørnehøj K., Balinda S.N., Muwanika V.B., Ademun Okurut A.R., Belsham G.J., Normann P., Siegismund H.R. and Alexandersen S. (2010) The role of African buffalos (Syncerus caffer) in the maintenance of foot-and-mouth disease in Uganda. *BMC Veterinary Research*, **6**, 54.
5. Ayelet G., Mahapatra M., Gelaye E., Egziabher B.G., Rufeal T., Sahle M., Ferris N.P., Wadsworth J., Hutchings G.H. and Knowles N.J. (2009) Genetic characterization of foot-and-mouth disease viruses, Ethiopia, 1981-2007. *Emerging Infectious Diseases*, **15** (9), 1409–1417.

6. Baele G., Li W.L.S., Drummond A.J., Suchard M.A. and Lemey P. (2013) Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, **30** (2), 239–243.

7. Balinda S.N., Sangula A.K., Heller R., Muwanika V.B., Belsham G.J., Masembe C. and Siegismund H.R. (2010) Diversity and transboundary mobility of serotype O foot-and-mouth disease virus in East Africa: implications for vaccination policies. *Infection, Genetics and Evolution*, **10** (7), 1058–1065.

8. Bastos A.D.S., Boshoff C.I., Keet D.F., Bengis R.G. and Thomson G.R. (2000) Natural transmission of foot-and-mouth disease virus between African buffalo (Syncerus caffer) and impala (Aepyceros melampus) in the Kruger National Park, South Africa. *Epidemiology and Infection*, **124** (3), 591–598.

9. Bastos A.D.S., Haydon D.T., Sangaré O., Boshoff C.I., Edrich J.L. and Thomson G.R. (2003) The implications of virus diversity within the SAT 2 serotype for control of foot-and-mouth disease in sub-Saharan Africa. *Journal of General Virology*, **84** (6), 1595–1606.

10. Bataille A., van der Meer F., Stegeman A. and Koch G. (2011) Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLOS Pathogens*, **7** (6), e1002094.

11. Beck E. and Strohmaier K. (1987) Subtyping of European foot-and-mouth disease virus strains by nucleotide sequence determination. *Journal of Virology*, **61** (5), 1621–1629.

12. Benson D.A., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J. and Sayers E.W. (2014) GenBank. *Nucleic Acids Research*, **42** (D1), D32–D37.

13. Bielejec F., Lemey P., Baele G., Rambaut A. and Suchard M.A. (2014) Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic Biology*, **63** (4), 493–504.

14. Bielejec F., Lemey P., Carvalho L.M., Baele G., Rambaut A. and Suchard M.A. (2014) πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics*, **15** (1), 133.

15. Boender G.J., Hagenaars T.J., Bouma A., Nodelijk G., Elbers A.R.W., de Jong M.C.M. and van Boven M. (2007) Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLOS Computational Biology*, **3** (4), e71.

16. Bogner P., Capua I., Lipman D.J., Cox N.J., et al. (2006) A global initiative on sharing avian flu data. *Nature*, **442** (7106), 981–981.

17. Bos M.E., Boven M.V., Nielen M., Bouma A., Elbers A.R., Nodelijk G., Koch G., Stegeman A. and Jong M.C.D. (2007) Estimating the day of highly pathogenic avian influenza (H7N7) virus introduction into a poultry flock based on mortality data. *Veterinary Research*, **38** (3), 12.

18. Bougnoux ME., Morand S. and d'Enfert C. (2002) Usefulness of multilocus sequence typing for characterization of clinical isolates of candida albicans. *Journal of Clinical Microbiology*, **40** (4), 1290–1297.

19. Bronsvoort B.M.d.C., Radford A.D., Tanya V.N., Nfon C., Kitching R.P. and Morgan K.L. (2004) Molecular epidemiology of foot-and-mouth disease viruses in the Adamawa province of Cameroon. *Journal of Clinical Microbiology*, **42** (5), 2186–2196.

20. Bronsvoort B.M.d.C., Tanya V.N., Kitching R.P., Nfon C., Hamman S.M. and Morgan K.L. (2003) Foot and mouth disease and livestock husbandry practices in the Adamawa Province of Cameroon. *Tropical Animal Health and Production*, **35** (6), 491–507.

21. Bull R.A., Luciani F., McElroy K., Gaudieri S., Pham S.T., Chopra A., Cameron B., Maher L., Dore G.J., White P.A. and Lloyd A.R. (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLOS Pathogens*, **7** (9), e1002243.

22. Carrillo C., Tulman E.R., Delhon G., Lu Z., Carreno A., Vagnozzi A., Kutish G.F. and Rock D.L. (2005) Comparative genomics of foot-and-mouth disease virus. *Journal of Virology*, **79** (10), 6487–6504.

23. Chikhi L., Sousa V.C., Luisi P., Goossens B. and Beaumont M.A. (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, **186** (3), 983–995.

24. Christensen L.S., Okurut R., Tjornehoj K., Normann P., Soerensen K. and Esau M. Characterisation of a new type O lineage of FMDV from Uganda with atypical clinical manifestations in domestic cattle. In: *Proceedings of the Open Session of the EUFMD Research Group*. Food and Agriculature Organisation of the United Nations. Rome, 2004, 159–162.

25. Christensen L.S., Normann P., Thykier-Nielsen S., Sørensen J.H., Stricker K.d. and Rosenørn S. (2005) Analysis of the epidemiological dynamics during the 1982–1983 epidemic of foot-and-mouth disease in Denmark based on molecular high-resolution strain identification. *Journal of General Virology*, **86** (9), 2577–2584.

26. Cottam E.M., Thébaud G., Wadsworth J., Gloster J., Mansley L., Paton D.J., King D.P. and Haydon D.T. (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, **275** (1637), 887–895.

27. Cottam E.M., Wadsworth J., Shaw A.E., Rowlands R.J., Goatley L., Maan S., Maan N.S., Mertens P.P.C., Ebert K., Li Y., Ryan E.D., Juleff N., Ferris N.P., Wilesmith J.W., Haydon D.T., King D.P., Paton D.J. and Knowles N.J. (2008) Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLOS Pathogens*, **4** (4), e1000050.

28. de Carvalho L.M.F., Santos L.B.L., Faria N.R. and de Castro Silveira W. (2013) Phylogeography of foot-and-mouth disease virus serotype O in Ecuador. *Infection, Genetics and Evolution*, **13**, 76–88.

29. de Carvalho L.M., Faria N.R., Perez A.M., Suchard M.A., Lemey P., de Castro Silveira W., Rambaut A. and Baele G. (2015) Spatio-temporal dynamics of foot-and-mouth disease virus in South America. *arXiv:1505.01105 [q-bio]*.

30. De Maio N., Wu CH., O'Reilly K.M. and Wilson D. (2015) New routes to phylogeography: a Bayesian structured coalescent approximation. *PLOS Genetics*, **11** (8), e1005421.

31. de Silva E., Ferguson N.M. and Fraser C. (2012) Inferring pandemic growth rates from sequence data. *Journal of The Royal Society Interface*, **9** (73), 1797–1808.

32. Deardon R., Brooks S.P., Grenfell B.T., Keeling M.J., Tildesley M.J., Savill N.J., Shaw D.J. and Woolhouse M.E.J. (2010) Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, **20** (1), 239–261.

33. Di Nardo A., Knowles N.J. and Paton D.J. (2011) Combining livestock trade patterns with phylogenetics to help understand the spread of foot and mouth disease in sub-Saharan Africa, the Middle East and Southeast Asia. *Revue Scientifique et Technique (International Office of Epizootics)*, **30** (1), 63–85.

34. Di Nardo A., Knowles N.J., Wadsworth J., Haydon D.T. and King D.P. (2014) Phylodynamic reconstruction of O CATHAY topotype foot-and-mouth disease virus epidemics in the Philippines. *Veterinary Research*, **45** (1), 90.

35. Didelot X., Gardy J. and Colijn C. (2014) Bayesian inference of infectious disease transmission from whole genome sequence data. *Molecular Biology and Evolution*, **31** (7), 1869–1879.

36. Drummond A.J., Rambaut A., Shapiro B. and Pybus O.G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22** (5), 1185–1192.

37. Drummond A.J., Ho S.Y.W., Phillips M.J. and Rambaut A. (2006) Relaxed phylogenetics and dating with confidence. *PLOS Biology*, **4** (5), e88.

38. Drummond A.J., Nicholls G.K., Rodrigo A.G. and Solomon W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161** (3), 1307–1320.

39. Drummond A.J., Suchard M.A., Xie D. and Rambaut A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29** (8), 1969–1973.

40. Drummond A. and Suchard M. (2010) Bayesian random local clocks, or one rate to rule them all. *BMC Bioinformatics*, **8**, 114.

41. Duchêne S., Holmes E.C. and Ho S.Y.W. (2014) Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B: Biological Sciences*, **281** (1786), 20140732.

42. Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32** (5), 1792–1797.

43. Elbers A.R.W., Fabri T.H.F., de Vries T.S., de Wit J.J., Pijpers A. and Koch G. (2004) The highly pathogenic avian influenza A (H7N7) virus epidemic in the Netherlands in 2003: lessons learned from the first five outbreaks. *Avian Diseases*, **48** (3), 691–705.

44. Famulare M. and Hu H. (2015) Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. *International Health*, **7** (2), 130–138.

45. Faria N.R., Suchard M.A., Rambaut A., Streicker D.G. and Lemey P. (2013) Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368** (1614), 20120196.

46. Felsenstein J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17** (6), 368–376.

47. Frost S.D.W., Pybus O.G., Gog J.R., Viboud C., Bonhoeffer S. and Bedford T. (2015) Eight challenges in phylodynamic inference. *Epidemics*, **10**, 88–92.

48. Frost S.D.W. and Volz E.M. (2010) Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365** (1548), 1879–1890.

49. Galecki A. and Burzykowski T. *Linear mixed-effects models using R: a step-by-step approach*. English. 2013 edition. New York, NY: Springer, Feb. 2013.

50. Gardy J.L., Johnston J.C., Ho Sui S.J., Cook V.J., Shah L., Brodkin E., Rempel S., Moore R., Zhao Y., Holt R., Varhol R., Birol I., Lem M., Sharma M.K., Elwood K., Jones S.J.M., Brinkman F.S.L., Brunham R.C. and Tang P. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England Journal of Medicine*, **364** (8), 730–739.

51. Gibson G.J. and Austin E.J. (1996) Fitting and testing spatio-temporal stochastic models with application in plant epidemiology. *Plant Pathology*, **45** (2), 172–184.

52. Gill M.S., Lemey P., Faria N.R., Rambaut A., Shapiro B. and Suchard M.A. (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, **30** (3), 713–724.

53. Gordo I. and Campos P.R. (2007) Patterns of genetic variation in populations of infectious agents. *BMC Evolutionary Biology*, **7**, 116.

54. Grazioli S., Moretti S., Barbieri I., Crosatti M. and Brocchi E. Use of monoclonal antibodies to identify and map new antigenic determinants involved in neutralisation on FMD viruses type SAT 1 and SAT 2. In: *Report of the Session of the Research Foot and Mouth Disease Group of the Standing Committee of the European Commission of Foot and Mouth Disease*. Food and Agriculature Organisation of the United Nations. Rome, 2006, 287–297.

55. Green P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82** (4), 711–732.

56. Grenfell B.T., Pybus O.G., Gog J.R., Wood J.L.N., Daly J.M., Mumford J.A. and Holmes E.C. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303** (5656), 327–332.

57. Griffiths R.C. and Tavare S. (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **344** (1310), 403–410.

58. Habiela M., Ferris N.P., Hutchings G.H., Wadsworth J., Reid S.M., Madi M., Ebert K., Sumption K.J., Knowles N.J., King D.P. and Paton D.J. (2010) Molecular characterization of foot-and-mouth disease viruses collected from Sudan. *Transboundary and Emerging Diseases*, **57** (5), 305–314.

59. Hall M.D., Knowles N.J., Wadsworth J., Rambaut A. and Woolhouse M.E.J. (2013) Reconstructing geographical movements and host species transitions of foot-and-mouth disease virus serotype SAT 2. *mBio*, **4** (5), e00591–13.

60. Hall M. and Rambaut A. (2014) Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions. *arXiv:1406.0428*. arXiv: 1406.0428.

61. Hasegawa M., Kishino H. and Yano Ta. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22** (2), 160–174.

62. Heled J. and Drummond A.J. (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27** (3), 570–580.

63. Heled J. and Drummond A.J. (2012) Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology*, **61** (1), 138–149.

64. Heller R., Chikhi L. and Siegismund H.R. (2013) The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLOS ONE*, **8** (5), e62992.

65. Hemadri D., Tosh C., Sanyal A. and Venkataramanan R. (2002) Emergence of a new strain of type of foot-and-mouth disease virus: its phylogenetic and evolutionary relationship with the PanAsia pandemic strain. *Virus Genes*, **25** (1), 23–34.

66. Ho S.Y.W., Phillips M.J., Cooper A. and Drummond A.J. (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, **22** (7), 1561–1568.

67. Hohna S., Defoin-Platel M. and Drummond A. Clock-constrained tree proposal operators in Bayesian phylogenetic inference. In: *8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008*. Oct. 2008, 1–7.

68. Holmes E.C. (2004) The phylogeography of human viruses. *Molecular Ecology*, **13** (4), 745–756.

69. Horn S., Prost S., Stiller M., Makowiecki D., Kuznetsova T., Benecke N., Pucher E., Hufthammer A.K., Schouwenburg C., Shapiro B. and Hofreiter M. (2014) Ancient mitochondrial DNA and the genetic history of Eurasian beaver (Castor fiber) in Europe. *Molecular Ecology*, **23** (7), 1717–1729.

70. Hughes J., Allen R.C., Baguelin M., Hampson K., Baillie G.J., Elton D., Newton J.R., Kellam P., Wood J.L.N., Holmes E.C. and Murcia P.R. (2012) Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLOS Pathogens*, **8** (12), e1003081.

71. Hunter P. (1998) Vaccination as a means of control of foot-and-mouth disease in sub-saharan Africa. *Vaccine*, **16** (2–3), 261–264.

72. Jackson A.L., O'Neill H., Maree F., Blignaut B., Carrillo C., Rodriguez L. and Haydon D.T. (2007) Mosaic structure of foot-and-mouth disease virus genomes. *Journal of General Virology*, **88** (Pt 2), 487–492.

73. Jagielski T., Augustynowicz-Kopeć E., Zozio T., Rastogi N. and Zwolska Z. (2010) Spoligotype-based comparative population structure analysis of multidrug-resistant and isoniazid-monoresistant Mycobacterium tuberculosis complex clinical isolates in Poland. *Journal of Clinical Microbiology*, **48** (11), 3899–3909.

74. Jamal S.M., Ferrari G., Ahmed S., Normann P. and Belsham G.J. (2011) Genetic diversity of foot-and-mouth disease virus serotype O in Pakistan and Afghanistan, 1997–2009. *Infection, Genetics and Evolution*, **11** (6), 1229–1238.

75. Jombart T., Eggo R.M., Dodd P.J. and Balloux F. (2011) Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, **106** (2), 383–390.

76. Jombart T., Cori A., Didelot X., Cauchemez S., Fraser C. and Ferguson N. (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLOS Computational Biology*, **10** (1), e1003457.

77. Jonges M., Bataille A., Enserink R., Meijer A., Fouchier R.A.M., Stegeman A., Koch G. and Koopmans M. (2011) Comparative analysis of avian influenza virus diversity in poultry and humans during a highly pathogenic avian influenza A (H7N7) virus outbreak. *Journal of Virology*, **85** (20), 10598–10604.

78. Kandeil A., El-Shesheny R., Kayali G., Moatasim Y., Bagato O., Darwish M., Gaffar A., Younes A., Farag T., Kutkat M.A. and Ali M.A. (2013) Characterization of the recent outbreak of foot-and-mouth disease virus serotype SAT2 in Egypt. *Archives of Virology*, **158** (3), 619–627.

79. Katoh K., Misawa K., Kuma Ki. and Miyata T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30** (14), 3059–3066.

80. Keele B.F., Giorgi E.E., Salazar-Gonzalez J.F., Decker J.M., Pham K.T., Salazar M.G., Sun C., Grayson T., Wang S., Li H., Wei X., Jiang C., Kirchherr J.L., Gao F., Anderson J.A., Ping LH., Swanstrom R., Tomaras G.D., Blattner W.A., Goepfert P.A., Kilby J.M., Saag M.S., Delwart E.L., Busch M.P., Cohen M.S., Montefiori D.C., Haynes B.F., Gaschen B., Athreya G.S., Lee H.Y., Wood N., Seoighe C., Perelson A.S., Bhattacharya T., Korber B.T., Hahn B.H. and Shaw G.M. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, **105** (21), 7552–7557.

81. Keeling M.J., Woolhouse M.E.J., Shaw D.J., Matthews L., Chase-Topping M., Haydon D.T., Cornell S.J., Kappey J., Wilesmith J. and Grenfell B.T. (2001) Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, **294** (5543), 813–817.

82. Khounsy S., Conlan J.V., Gleeson L.J., Westbury H.A., Colling A., Paton D.J., Ferris N.P., Valarcher JF., Wadsworth J., Knowles N.J. and Blacksell S.D. (2009) Molecular epidemiology of foot-and-mouth disease viruses from South East Asia 1998-2006: the Lao perspective. *Veterinary Microbiology*, **137** (1-2), 178–183.

83. Knowles N.J., Nazem Shirazi M.H., Wadsworth J., Swabey K.G., Stirling J.M., Statham R.J., Li Y., Hutchings G.H., Ferris N.P., Parlak Ü., Özyörük F., Sumption K.J., King D.P. and Paton D.J. (2009) Recent spread of a new strain (A-Iran-05) of foot-and-mouth disease virus type A in the Middle East. *Transboundary and Emerging Diseases*, **56** (5), 157–169.

84. Knowles N.J., Davies P.R., Midgley R.J. and Valarcher JF. Identification of a ninth foot-and-mouth disease virus type O topotype and evidence for a recombination event in its evolution. In: *Report of the Session of the Research Group of the Standing Technical Committee of EUFMD;* Chania, Crete, Greece, 2004.

85. Knowles N.J., Samuel A.R., Davies P.R., Midgley R.J. and Valarcher JF. (2005) Pandemic strain of foot-and-mouth disease virus serotype O. *Emerging Infectious Diseases*, **11** (12), 1887–1893.

86. Knowles N.J., Wadsworth J., Reid S.M., Swabey K.G., El-Kholy A.A., Abd El-Rahman A.O., Soliman H.M., Ebert K., Ferris N.P., Hutchings G.H., Statham R.J., King D.P. and Paton D.J. (2007) Foot-and-mouth disease virus serotype A in Egypt. *Emerging Infectious Diseases*, **13** (10), 1593–1596.

87. Knowles N. and Samuel A. (2003) Molecular epidemiology of foot-and-mouth disease virus. *Virus Research*, **91** (1), 65–80.

88. Koelle K., Cobey S., Grenfell B. and Pascual M. (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, **314** (5807), 1898–1903.

89. König G., Palma E., Maradei E. and Piccone M. (2007) Molecular epidemiology of foot-and-mouth disease virus types A and O isolated in Argentina during the 2000–2002 epizootic. *Veterinary Microbiology*, **124** (1–2), 1–15.

90. König G., Blanco C., Knowles N.J., Palma E.L., Maradei E. and Piccone M.E. (2001) Phylogenetic analysis of foot-and-mouth disease viruses isolated in Argentina. *Virus Genes*, **23** (2), 175–181.

91. Kühnert D., Wu CH. and Drummond A.J. (2011) Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, **11** (8), 1825–1841.

92. Leigh Brown A.J., Lycett S.J., Weinert L., Hughes G.J., Fearnhill E. and Dunn D.T. (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Diseases*, **204** (9), 1463–1469.

93. Lemey P., Rambaut A., Bedford T., Faria N., Bielejec F., Baele G., Russell C.A., Smith D.J., Pybus O.G., Brockmann D. and Suchard M.A. (2014) Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLOS Pathogens*, **10** (2), e1003932.

94. Lemey P., Rambaut A., Drummond A.J. and Suchard M.A. (2009) Bayesian phylogeography finds its roots. *PLOS Computational Biology*, **5** (9), e1000520.

95. Lemey P., Rambaut A., Welch J.J. and Suchard M.A. (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, **27** (8), 1877–1885.

96. Lewis F., Hughes G.J., Rambaut A., Pozniak A. and Leigh Brown A.J. (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLOS Medicine*, **5** (3), e50.

97. Lewis-Rogers N., McClellan D.A. and Crandall K.A. (2008) The evolution of foot-and-mouth disease virus: impacts of recombination and selection. *Infection, Genetics and Evolution*, **8** (6), 786–798.

98. Lin JH., Chiu SC., Cheng JC., Chang HW., Hsiao KL., Lin YC., Wu HS., Salemi M. and Liu HF. (2011) Phylodynamics and molecular evolution of influenza A virus nucleoprotein genes in Taiwan between 1979 and 2009. *PLOS ONE*, **6** (8), e23454.

99. Malirat V., de Barros J.J.F., Bergmann I.E., Campos R.d.M., Neitzert E., da Costa E.V., da Silva E.E., Falczuk A.J., Pinheiro D.S.B., de Vergara N., Cirvera J.L.Q., Maradei E. and Di Landro R. (2007) Phylogenetic analysis of foot-and-mouth disease virus type O re-emerging in free areas of South America. *Virus Research*, **124** (1–2), 22–28.

100. Matsumura S., Inoshima Y. and Ishiguro N. (2014) Reconstructing the colonization history of lost wolf lineages by the analysis of the mitochondrial genome. *Molecular Phylogenetics and Evolution*, **80**, 105–112.

101. Metzker M.L. (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11** (1), 31–46.

102. Minin V.N., Bloomquist E.W. and Suchard M.A. (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, **25** (7), 1459–1471.

103. Minin V.N. and Suchard M.A. (2008) Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*, **56** (3), 391–412.

104. Mollentze N., Nel L.H., Townsend S., Roux K.l., Hampson K., Haydon D.T. and Soubeyrand S. (2014) A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, **281** (1782), 20133251.

105. Morelli M.J., Thébaud G., Chadœuf J., King D.P., Haydon D.T. and Soubeyrand S. (2012) A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLOS Computational Biology*, **8** (11), e1002768.

106. Notohara M. (1990) The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29** (1), 59–75.

107. Numminen E., Chewapreecha C., Sirén J., Turner C., Turner P., Bentley S.D. and Corander J. (2014) Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society B: Biological Sciences*, **281** (1794), 20141324.

108. Opperman P.A., Maree F.F., Van Wyngaardt W., Vosloo W. and Theron J. (2012) Mapping of antigenic determinants on a SAT2 foot-and-mouth disease virus using chicken single-chain antibody fragments. *Virus Research*, **167** (2), 370–379.

109. Padhi A. and Ma L. (2014) Genetic and epidemiological insights into the emergence of peste des petits ruminants virus (PPRV) across Asia and Africa. *Scientific Reports*, **4**, 7040.

110. Phillips J.E., Stallknecht D.E., Perkins T.A., McClure N.S. and Mead D.G. (2014) Evolutionary dynamics of West Nile virus in Georgia, 2001–2011. *Virus Genes*, **49** (1), 132–136.

111. Phologane B.S., Dwarka R.M., Haydon D.T., Gerber L.J. and Vosloo W. (2008) Molecular characterization of SAT-2 foot-and-mouth disease virus isolates obtained from cattle during a four-month period in 2001 in Limpopo Province, South Africa. *The Onderstepoort Journal of Veterinary Research*, **75** (4), 267–277.

112. Pond K., L S., Posada D., Gravenor M.B., Woelk C.H. and Frost S.D.W. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*, **23** (10), 1891–1901.

113. Pond S.L.K. and Frost S.D.W. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, **22** (5), 1208–1222.

114. Pybus O.G., Charleston M.A., Gupta S., Rambaut A., Holmes E.C. and Harvey P.H. (2001) The epidemic behavior of the hepatitis C virus. *Science*, **292** (5525), 2323–2325.

115. Pybus O.G. and Rambaut A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, **10** (8), 540–550.

116. Pybus O.G., Rambaut A. and Harvey P.H. (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155** (3), 1429–1437.

117. Rasmussen D.A., Ratmann O. and Koelle K. (2011) Inference for non-linear epidemiological models using genealogies and time series. *PLOS Computational Biology*, **7** (8), e1002136.

118. Reeve R., Blignaut B., Esterhuysen J.J., Opperman P., Matthews L., Fry E.E., de Beer T.A.P., Theron J., Rieder E., Vosloo W., O'Neill H.G., Haydon D.T. and Maree F.F. (2010) Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. *PLOS Computational Biology*, **6** (12), e1001027.

119. Rodrigo A. The coalescent: population genetic inference using genealogies. In: *The phylogenetics handbook*. 2nd edition. Cambridge University Press, 2011, 551–563.

120. Rodriguez F., Oliver J.L., Marın A. and Medina J.R. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, **142** (4), 485–501.

121. Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A. and Huelsenbeck J.P. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61** (3), 539–542.

122. Rweyemamu M., Roeder P., Mackay D., Sumption K., Brownlie J., Leforban Y., Valarcher J.F., Knowles N.J. and Saraiva V. (2008) Epidemiological patterns of foot-and-mouth disease worldwide. *Transboundary and Emerging Diseases*, **55** (1), 57–72.

123. Sahle M., Dwarka R.M., Venter E.H. and Vosloo W. (2007) Study of the genetic heterogeneity of SAT-2 foot-and-mouth disease virus in sub-Saharan Africa with specific focus on East Africa. *The Onderstepoort Journal of Veterinary Research*, **74** (4), 289–299.

124. Salipante S.J. and Hall B.G. (2011) Inadequacies of minimum spanning trees in molecular epidemiology. *Journal of Clinical Microbiology*, **49** (10), 3568–3575.

125. Samuel A.R. and Knowles N.J. (2001) Foot-and-mouth disease type O viruses exhibit genetically and geographically distinct evolutionary lineages (topotypes). *Journal of General Virology*, **82** (3), 609–621.

126. Samuel A.R., Knowles N.J. and Mackay D.K.J. (1999) Genetic analysis of type O viruses responsible for epidemics of foot-and-mouth disease in North Africa. *Epidemiology & Infection*, **122** (03), 529–538.

127. Sangaré O., Bastos A.D.S., Venter E.H. and Vosloo W. (2004) A first molecular epidemiological study of SAT-2 type foot-and-mouth disease viruses in West Africa. *Epidemiology and Infection*, **132** (3), 525–532.

128. Sangula A.K., Siegismund H.R., Belsham G.J., Balinda S.N., Masembe C. and Muwanika V.B. (2011) Low diversity of foot-and-mouth disease serotype C virus in Kenya: evidence for probable vaccine strain re-introductions in the field. *Epidemiology and Infection*, **139** (02), 189–196.

129. Sangula A., Belsham G., Muwanika V., Heller R., Balinda S. and Siegismund H. (2010) Co-circulation of two extremely divergent serotype SAT 2 lineages in Kenya highlights challenges to foot-and-mouth disease control. *Archives of Virology*, **155** (10), 1625–1630.

130. Schmid F. and Schmidt A. (2006) Nonparametric estimation of the coefficient of overlapping—theory and empirical application. *Computational Statistics and Data Analysis*, **50** (6), 1583–1596.

131. Schumann K.R., Knowles N.J., Davies P.R., Midgley R.J., Valarcher JF., Raoufi A.Q., McKenna T.S., Hurtle W., Burans J.P., Martin B.M., Rodriguez L.L. and Beckham T.R. (2008) Genetic characterization and molecular epidemiology of foot-and-mouth disease viruses isolated from Afghanistan in 2003–2005. *Virus Genes*, **36** (2), 401–413.

132. Shapiro B., Rambaut A. and Drummond A.J. (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, **23** (1), 7–9.

133. Sheather S.J. and Jones M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53** (3), 683–690.

134. Slatkin M. and Hudson R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129** (2), 555–562.

135. Sobhy N.M., Mor S.K., Mohammed M.E.M., Bastawecy I.M., Fakhry H.M., Youssef C.R. and Goyal S.M. (2014) Phylogenetic analysis of Egyptian foot and mouth disease virus endemic strains. *Journal of American Science*, **10** (9), 133–138.

136. Spada E., Sagliocca L., Sourdis J., Garbuglia A.R., Poggi V., Fusco C.D. and Mele A. (2004) Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *Journal of Clinical Microbiology*, **42** (9), 4230–4236.

137. Stack J.C., Welch J.D., Ferrari M.J., Shapiro B.U. and Grenfell B.T. (2010) Protocols for sampling viral sequences to study epidemic dynamics. *Journal of the Royal Society Interface*, **7** (48), 1119–1127.

138. Stadler T. (2010) Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, **267** (3), 396–404.

139. Stadler T., Kouyos R., von Wyl V., Yerly S., Böoni J., Bürgisser P., Klimkait T., Joos B., Rieder P., Xie D., Günthard H.F., Drummond A.J., Bonhoeffer S. and Swiss HIV Cohort Study the. (2011) Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, **29** (1), 347–357.

140. Stegeman A., Bouma A., Elbers A.R.W., De Jong M.C.M., Nodelijk G., De Klerk F., Koch G. and Van Boven M. (2004) Avian influenza A virus (H7N7) epidemic in the Netherlands in 2003: course of the epidemic and effectiveness of control measures. *Journal of Infectious Diseases*, **190** (12), 2088–2095.

141. Stram Y., Engel O., Rubinstein M., Kuznetzova L., Balaish M., Yadin H., Istumin S. and Gelman B. (2011) Multiple invasions of O1 FMDV serotype into Israel revealed by genetic analysis of VP1 genes of Israeli's isolates from 1989 to 2007. *Veterinary Microbiology*, **147** (3–4), 398–402.

142. Tamura K. and Nei M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, **10** (3), 512–526.

143. Thalmann O., Shapiro B., Cui P., Schuenemann V.J., Sawyer S.K., Greenfield D.L., Germonpré M.B., Sablin M.V., López-Giráldez F., Domingo-Roura X., Napierala H., Uerpmann HP., Loponte D.M., Acosta A.A., Giemsch L., Schmitz R.W., Worthington B., Buikstra J.E., Druzhkova A., Graphodatsky A.S., Ovodov N.D., Wahlberg N., Freedman A.H., Schweizer R.M., Koepfli KP., Leonard J.A., Meyer M., Krause J., Pääbo S., Green R.E. and Wayne R.K. (2013) Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science*, **342** (6160), 871–874.

144. Thomson G., Vosloo W. and Bastos A. (2003) Foot and mouth disease in wildlife. *Virus Research*, **91** (1), 145–161.

145. Tully D.C. and Fares M.A. (2008) The tale of a modern animal plague: tracing the evolutionary history and determining the time-scale for foot and mouth disease virus. *Virology*, **382** (2), 250–256.

146. Valarcher J.F., Leforban Y., Rweyemamu M., Roeder P.L., Gerbier G., Mackay D.K.J., Sumption K.J., Paton D.J. and Knowles N.J. (2008) Incursions of foot-and-mouth disease virus into Europe between 1985 and 2006. *Transboundary and Emerging Diseases*, **55** (1), 14–34.

147. Valdazo-González B., Knowles N.J., Hammond J. and King D.P. (2012) Genome sequences of SAT 2 foot-and-mouth disease viruses from Egypt and Palestinian Autonomous Territories (Gaza Strip). *Journal of Virology*, **86** (16), 8901–8902.

148. Van Borm S., Jonges M., Lambrecht B., Koch G., Houdart P. and Berg T. van den. (2014) Molecular epidemiological analysis of the transboundary transmission of 2003 highly pathogenic avian influenza H7N7 outbreaks between the Netherlands and Belgium. *Transboundary and Emerging Diseases*, **61** (1), 86–90.

149. van Rensburg H.G., Henry T.M. and Mason P.W. (2004) Studies of genetically defined chimeras of a European type A virus and a South African Territories type 2 virus reveal growth determinants for foot-and-mouth disease virus. *Journal of General Virology*, **85** (Pt 1), 61–68.

150. van Rensburg H.G. and Nel L.H. (1999) Characterization of the structural-protein-coding region of SAT 2 type foot-and-mouth disease virus. *Virus Genes*, **19** (3), 229–233.

151. van Rensburg H., Haydon D., Joubert F., Bastos A., Heath L. and Nel L. (2002) Genetic heterogeneity in the foot-and-mouth disease virus Leader and 3C proteinases. *Gene*, **289** (1-2), 19–29.

152. Vaughan T.G., Kühnert D., Popinga A., Welch D. and Drummond A.J. (2014) Efficient Bayesian inference under the structured coalescent. *Bioinformatics*, 2272–2279.

153. Volz E.M. (2012) Complex population dynamics and the coalescent under neutrality. *Genetics*, **190** (1), 187–201.

154. Volz E.M. and Frost S.D.W. (2013) Inferring the source of transmission with phylogenetic data. *PLOS Computational Biology*, **9** (12), e1003397.

155. Volz E.M., Pond S.L.K., Ward M.J., Brown A.J.L. and Frost S.D.W. (2009) Phylodynamics of infectious disease epidemics. *Genetics*, **183** (4), 1421–1430.

156. Vosloo W., Bastos A.D.S., Sangare O., Hargreaves S.K. and Thomson G.R. (2002) Review of the status and control of foot and mouth disease in sub-Saharan Africa. *Revue Scientifique et Technique (International Office of Epizootics)*, **21** (3), 437–449.

157. Vosloo W., Thompson P.N., Botha B., Bengis R.G. and Thomson G.R. (2009) Longitudinal study to investigate the role of impala (Aepyceros melampus) in foot-and-mouth disease maintenance in the Kruger National Park, South Africa. *Transboundary and Emerging Diseases*, **56** (1-2), 18–30.

158. Vosloo W., Boshoff K., Dwarka R. and Bastos A. (2002) The possible role that buffalo played in the recent outbreaks of foot-and-mouth disease in South Africa. *Annals of the New York Academy of Sciences*, **969** (1), 187–190.

159. Vrancken B., Rambaut A., Suchard M.A., Drummond A., Baele G., Derdelinckx I., Van Wijngaerden E., Vandamme AM., Van Laethem K. and Lemey P. (2014) The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLOS Computational Biology*, **10** (4), e1003505.

160. Waheed U., Parida S., Khan Q.M., Hussain M., Ebert K., Wadsworth J., Reid S.M., Hutchings G.H., Mahapatra M., King D.P., Paton D.J. and Knowles N.J. (2011) Molecular characterisation of foot-and-mouth disease viruses from Pakistan, 2005–2008. *Transboundary and Emerging Diseases*, **58** (2), 166–172.

161. Wakeley J. *Coalescent theory: an introduction*. English. New title edition. Greenwood Village, Colo: Roberts & Co, Aug. 2008.

162. Wallinga J. and Lipsitch M. (2007) How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, **274** (1609), 599–604.

163. Welch D., Nicholls G.K., Rodrigo A. and Solomon W. (2005) Integrating genealogy and epidemiology: the ancestral infection and selection graph as a model for reconstructing host virus histories. *Theoretical Population Biology*, **68** (1), 65–75.

164. Whidden C., Zeh N. and Beiko R.G. (2014) Supertrees based on the subtree prune-and-regraft distance. *Systematic Biology*, **63** (4), 566–581.

165. White L.F. and Pagano M. (2008) A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, **27** (16), 2999–3016.

166. Wilson I.J. and Balding D.J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150** (1), 499–510.

167. Worby C.J., Chang HH., Hanage W.P. and Lipsitch M. (2014) The distribution of pairwise genetic distances: a tool for investigating disease transmission. *Genetics*, **198** (4), 1395–1404.

168. Worby C.J., Lipsitch M. and Hanage W.P. (2014) Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLOS Computational Biology*, **10** (3), e1003549.

169. Wright C.F., Knowles N.J., Di Nardo A., Paton D.J., Haydon D.T. and King D.P. (2013) Reconstructing the origin and transmission dynamics of the 1967–68 foot-and-mouth disease epidemic in the United Kingdom. *Infection, Genetics and Evolution*, **20**, 230–238.

170. Yoon S.H., Park W., King D.P. and Kim H. (2011) Phylogenomics and molecular evolution of foot-and-mouth disease virus. *Molecules and Cells*, **31** (5), 413–421.

171. Ypma R.J.F., Bataille A.M.A., Stegeman A., Koch G., Wallinga J. and van Ballegooijen W.M. (2011) Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, **279** (1728), 444–450.

172. Ypma R.J.F., Jonges M., Bataille A., Stegeman A., Koch G., van Boven M., Koopmans M., van Ballegooijen W.M. and Wallinga J. (2013) Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *Journal of Infectious Diseases*, **207** (5), 730–735.

173. Ypma R.J.F., van Ballegooijen W.M. and Wallinga J. (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, **195** (3), 1055–1062.

Appendix A

Summary of sequences used in analyses of foot-and-mouth disease virus serotype SAT 2

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
ANG/04/74	Angola	1974	[9]	XI	AF479417, DQ009736	Not given
BAR/10/2012	Bahrain	2012	[2]	IV	JX570610	Cattle
BAR/12/2012	Bahrain	2012	[2]	IV	JX570611	Cattle
BAR/13/2012	Bahrain	2012	[2]	IV	JX570612	Cattle
BAR/16/2012	Bahrain	2012	[2]	IV	JX570613	Cattle
BAR/28/2012	Bahrain	2012	[2]	IV	JX570614	Cattle
BOT/3/77	Botswana	1977	[59]	III	KF112928	Cattle
BOT 08/78	Botswana	1978	[59]	III	KF112929	Cattle
BOT/1/98	Botswana	1998	[9]	II	AF367122	*S. caffer*
BOT/18/98	Botswana	1998	[9]	II	AF367123	*S. caffer*
BOT/29/98	Botswana	1998	[9]	III	AF367124	*S. caffer*
BOT/31/98	Botswana	1998	[9]	III	AF367125	*S. caffer*
BUN/4/86	Burundi	1986	[59]	VIII	KF112930	Not given
BUN/1/91	Burundi	1991	[9]	IV	AF367111	Cattle
CAR/P12/2000 (VDI 44/1)	Cameroon	2000	[58]	VII	HM211082	Cattle
CAR/1/2005	Cameroon	2005	[2]	VII	JX570615	Cattle
CAR/8/2005	Cameroon	2005	[2]	VII	JX570616	Cattle
IVY/3/90	Côte d'Ivoire	1990	[59]	V	KF112957	Cattle
ZAI/01/74	DR Congo	1974		VIII	DQ009737	Not given
ZAI/1/82	DR Congo	1982	[9]	X	AF367100	Cattle
EGY/10/2012	Egypt	2012	[2]	VII	JX570623	Cattle
EGY/11/2012	Egypt	2012	[2]	VII	JX570624	Cattle
EGY/13/2012	Egypt	2012	[2]	VII	JX570625	Cattle
EGY/14/2012	Egypt	2012	[2]	VII	JX570626	Cattle
EGY/15/2012	Egypt	2012	[2]	VII	JX570627	Cattle
EGY/2/2012	Egypt	2012	[2]	VII	JX570617	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
EGY/3/2012	Egypt	2012	[2]	VII	JX570618	Cattle
EGY/4/2012	Egypt	2012	[2]	VII	JX570619	Cattle
EGY/5/2012	Egypt	2012	[2]	VII	JX570620	Cattle
EGY/6/2012	Egypt	2012	[2]	VII	JX570621	Cattle
EGY/9/2012	Egypt	2012	[2]	VII	JX570622	Cattle
EGY/23/2012	Egypt	2012	[78]	VII	JX013980	Cattle
EGY/26/2012	Egypt	2012	[78]	VII	JX013979	*B. bubalis*
EGY/7/2012	Egypt	2012	[78]	VII	JX013978	*B. bubalis*
EGY/16/2012	Egypt	2012	[59]	VII	KF112931	*B. bubalis*
EGY/17/2012	Egypt	2012	[59]	VII	KF112932	*B. bubalis*
EGY/21/2012	Egypt	2012	[59]	VII	KF112933	Cattle
EGY/22/2012	Egypt	2012	[59]	VII	KF112934	Cattle
EGY/28/2012	Egypt	2012	[59]	VII	KF112935	Cattle
EGY/29/2012	Egypt	2012	[59]	VII	KF112936	Cattle
EGY/31/2012	Egypt	2012	[59]	VII	KF112937	Cattle
EGY/9/2012	Egypt	2012	[147]	VII	JX014255	Not given
ERI/12/98	Eritrea	1998	[9, 118]	VII	AF367126, GU194494	Cattle
ERI/1/98	Eritrea	1998	[123]	VII	AY343933	Cattle
ERI/4/98	Eritrea	1998	[123]	VII	AY343934	Cattle
ETH/1/91	Ethiopia	1989	[5, 123]	IV	FJ798158	Cattle
ETH/1/90	Ethiopia	1989	[123]	IV	AY343935	Cattle
ETH/2/90	Ethiopia	1989	[123]	IV	AY343936	Cattle
ETH/3/91	Ethiopia	1991	[5]	XIV	FJ798160	Cattle
ETH/2/91	Ethiopia	1991	[5, 123]	XIV	AY343938, FJ798159	Cattle
ETH/2/2007	Ethiopia	2007	[5]	XIII	FJ798161	Cattle
ETH/42/2009	Ethiopia	2009	[59]	XIII	KF112938	Not given

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
ETH/48/2009	Ethiopia	2009	[59]	XIII	KF112939	Not given
ETH/51/2009	Ethiopia	2009	[59]	XIII	KF112940	Cattle
ETH/52/2009	Ethiopia	2009	[59]	XIII	KF112941	Cattle
ETH/53/2009	Ethiopia	2009	[59]	XIII	KF112942	Cattle
ETH/56/2009	Ethiopia	2009	[59]	XIII	KF112943	Cattle
ETH/64/2009	Ethiopia	2009	[59]	XIII	KF112944	Cattle
ETH/65/2009	Ethiopia	2009	[59]	XIII	KF112945	Cattle
ETH/67/2009	Ethiopia	2009	[59]	XIII	KF112946	Cattle
ETH/68/2009	Ethiopia	2009	[59]	XIII	KF112947	Cattle
ETH/69/2009	Ethiopia	2009	[59]	XIII	KF112948	Cattle
ETH/70/2009	Ethiopia	2009	[59]	XIII	KF112949	Cattle
ETH/72/2009	Ethiopia	2009	[59]	XIII	KF112950	Pig
ETH/73/2009	Ethiopia	2009	[59]	XIII	KF112951	Cattle
ETH/74/2009	Ethiopia	2009	[59]	XIII	KF112952	Cattle
ETH/75/2009	Ethiopia	2009	[59]	XIII	KF112953	Cattle
ETH/76/2009	Ethiopia	2009	[59]	XIII	KF112954	Cattle
ETH/77/2009	Ethiopia	2009	[59]	XIII	KF112955	Cattle
ETH/2/2010	Ethiopia	2010	[59]	XIII	KF112956	Cattle
GAM/8/79	Gambia	1979	[9]	VI	AF479410	Cattle
GAM/9/79	Gambia	1979	[9]	VI	AF479411	Cattle
PAT/1/2012	Gaza Strip	2012	[2, 147]	VII	JX014256, JX570637	Cattle
GHA/2/90	Ghana	1990	[9]	V	AF479415	Not given
GHA/08/91	Ghana	1991	[9]	V	AF479416, DQ009732	Not given
KEN/3/57	Kenya	1957		IX	AJ251473	Not given
SAT2-3kenya 11/60	Kenya	1960	[22]	IX	AY593849, NC_003992	Not given
KEN/2/76	Kenya	1974	[123]	IV	AY343940	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
K81/81	Kenya	1981	[129]	IV	HM623678	Cattle
K46/82	Kenya	1982	[129]	IV	HM623679	Cattle
K65/82	Kenya	1982	[129]	IX	HM623680	Cattle
K151/83	Kenya	1983	[129]	IV	HM623683	Cattle
K70/83	Kenya	1983	[129]	IV	HM623681	Cattle
KEN/1/84	Kenya	1984	[123]	IV	AY344505	Cattle
KEN/2/84	Kenya	1984	[123]	IX	AY343941	Cattle
K34/84	Kenya	1984	[129]	IV	HM623684	Cattle
K52/84	Kenya	1984	[129]	IV	HM623685	Cattle
KEN/1/85	Kenya	1985	[123]	IX	AY343942	Cattle
KEN/1/86	Kenya	1986	[123]	IV	AY343943	Cattle
K37/86	Kenya	1986	[129]	IV	HM623686	Cattle
KEN/1/87	Kenya	1987	[123]	IV	AY343944	Cattle
KEN/2/87	Kenya	1987	[123]	IV	AY343945	Cattle
K13/87	Kenya	1987	[129]	IV	HM623687	Cattle
KEN/2/88	Kenya	1988	[123]	IX	AY343946	Cattle
KEN/1/89	Kenya	1989	[123]	IX	AY343947	Cattle
K40/90	Kenya	1990	[129]	IX	HM623688	Cattle
KEN/33/91	Kenya	1991	[123]	IV	AY343950	Cattle
KEN/28/91	Kenya	1991	[123]	IX	AY343948	Cattle
KEN/8/91	Kenya	1991	[123]	IX	AY343949	Cattle
K14/91	Kenya	1991	[129]	IX	HM623689	Cattle
KEN/1/92	Kenya	1992	[123]	IX	AY343953	Cattle
KEN/3/92	Kenya	1992	[123]	IX	AY343951	Cattle
KEN/6/92	Kenya	1992	[123]	IX	AY343952	Cattle
K32/92	Kenya	1992	[129]	IX	HM623690	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
K3/93	Kenya	1993	[129]	IX	HM623691	Cattle
KEN/1/94	Kenya	1994	[123]	IX	AY343954	Cattle
KEN/2/94	Kenya	1994	[123]	IX	AY343955	Cattle
K37/94	Kenya	1994	[129]	IV	HM623694	Cattle
K25/94	Kenya	1994	[129]	IX	HM623693	Cattle
K5/94	Kenya	1994	[129]	IX	HM623692	Cattle
KEN/3/95	Kenya	1995	[123]	IV	AY343957	Cattle
KEN/7/95	Kenya	1995	[123]	IX	AY343956	Cattle
K37/95	Kenya	1995	[129]	IX	HM623695	Cattle
K39/95	Kenya	1995	[129]	IX	HM623696	Cattle
KEN/1/96	Kenya	1996	[123]	IX	AY343960	Cattle
KEN/11/96	Kenya	1996	[123]	IX	AY343958	Cattle
KEN/7/96	Kenya	1996	[123]	IX	AY343959	Cattle
K77/96	Kenya	1996	[129]	IX	HM623697	Cattle
KEN/16/98	Kenya	1998	[123]	IV	AY343962	Cattle
KEN/7A/98	Kenya	1998	[123]	IX	AY343961	Cattle
KEN/5/99	Kenya	1999	[9]	IV	AF367131	Cattle
KEN/7/99	Kenya	1999	[9]	IV	AF367132	Cattle
KEN/9/99	Kenya	1999	[9]	IV	AF367133	Cattle
K49/99	Kenya	1999	[129]	IV	HM623698	Cattle
KEN/08/99	Kenya	1999		IV	DQ009729	Not given
K13/02	Kenya	2002	[129]	IV	HM623699	Cattle
KEN_002/2002	Kenya	2002		IV	JF749861	Not given
K120/04	Kenya	2004	[129]	IV	HM623700	Cattle
K70/05	Kenya	2005	[129]	IV	HM623701	Cattle
K6/06	Kenya	2006	[129]	IV	HM623702	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
K12/07	Kenya	2007	[129]	IV	HM623703	Cattle
K15/07	Kenya	2007	[129]	IV	HM623704	Cattle
K17/07	Kenya	2007	[129]	IV	HM623705	Cattle
K20/07	Kenya	2007	[129]	IV	HM623706	Cattle
K42/07	Kenya	2007	[129]	IV	HM623707	Cattle
K59/07	Kenya	2007	[129]	IV	HM623708	Cattle
K67/07	Kenya	2007	[129]	IV	HM623709	Cattle
KEN/11/2009	Kenya	2009	[2]	IV	JX570628	Cattle
KEN/122/2009	Kenya	2009	[2]	IV	JX570630	Cattle
KEN/13/2009	Kenya	2009	[2]	IV	JX570629	Cattle
LIB/1/2003	Libya	2003	[2]	VII	JX570631	Cattle
LIB/7/2003	Libya	2003	[2]	VII	JX570632	Cattle
LIB/39/2012	Libya	2012	[2]	VII	JX570633	Cattle
LIB/40/2012	Libya	2012	[2]	VII	JX570634	Cattle
LIB/41/2012	Libya	2012	[2]	VII	JX570635	Cattle
MAL/3/75	Malawi	1975	[9]	IV	AF367099	Not given
MOZ/1/70	Mozambique	1970	[59]	I	KF112959	Not given
MOZ/1/79	Mozambique	1979	[9]	I	AF367137	Not given
MOZ/4/83	Mozambique	1983	[9]	I	AF367101	Not given
SWA/4/89	Namibia	1989	[59]	III	KF112969	Cattle
NAM/286/98	Namibia	1998	[9]	II	AF367127	*S. caffer*
NAM/292/98	Namibia	1998	[9]	II	AF367128	*S. caffer*
NAM/304/98	Namibia	1998	[9]	II	AF367129	*S. caffer*
NGR/15/2005	Niger	2005	[59]	VII	KF112960	Cattle
NIG/2/75	Nigeria	1975	[9]	V	AF367139	Cattle
NIG/2/2007	Nigeria	2007	[2]	VII	JX570636	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
NYE/29/90	North Yemen	1990	[59]	IV	KF112961	Cattle
RWA/13/96	Rwanda	1996	[59]	VIII	KF112964	Not given
RWA/1/00	Rwanda	2000	[9]	VIII	AF367134	Cattle
RWA/02/01	Rwanda	2001		VIII	DQ009730	Not given
RWA/1/2004	Rwanda	2004	[59]	VIII	KF112963	Cattle
SAU/6/00	Saudi Arabia	2000	[9]	VII	AF367135, AY297948	Cattle
SEN/05/75	Senegal	1975	[9]	V	AF367140, DQ009738	Cattle
SEN/7/79	Senegal	1979	[9]	VI	AF479412	Not given
SEN/07/83	Senegal	1983	[9]	VI	AF479414, DQ009733	Not given
SEN/3/83	Senegal	1983	[9]	VI	AF479413	Not given
SEN/27/2009	Senegal	2009	[59]	VII	KF112967	Cattle
SAT2-2 106/67	South Africa	1959	[22]	I	AY593848	Not given
SA/2/67	South Africa	1967	[59]	I	KF112965	*A. melampus*
SAR/3/79	South Africa	1979	[59]	I	KF112966	Cattle
PAL/5/83	South Africa	1983	[9]	I	AF367102	Cattle
SAR/16/83	South Africa	1983		I	DQ009734	Not given
KNP/1/85	South Africa	1985	[59]	I	KF112958	*A. melampus*
KNP/16/88	South Africa	1988	[9]	I	AF367104	*A. melampus*
KNP/17/88	South Africa	1988	[9]	I	AF367105	*A. melampus*
KNP/18/88	South Africa	1988	[9]	I	AF367138	*A. melampus*
KNP/19/88	South Africa	1988	[9]	I	AF367106	*A. melampus*
KNP/20/88	South Africa	1988	[9]	I	AF367107	*A. melampus*
KNP/7/88	South Africa	1988	[9]	I	AF367103	Not given
KNP/19/89	South Africa	1989	[9]	I	AF367110, DQ009735	*S. caffer*
KNP/2/89	South Africa	1989	[9, 118]	I	AF367109, GU194488	*A. melampus*
KNP/183/91	South Africa	1991	[9]	I	AF367112	*S. caffer*

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
KNP/1/92	South Africa	1992	[9]	I	AF367114	*A. melampus*
KNP/32/92	South Africa	1992	[9]	I	AF367115	*S. caffer*
KNP/51/93	South Africa	1993	[118]	I	GU194489	*A. melampus*
KNP/18/95	South Africa	1995	[9]	I	AF367118	*S. caffer*
KNP/31/95	South Africa	1995	[9]	I	AF367119	*S. caffer*
SAR/1/01	South Africa	2001	[111]	I	AY442903	Not given
SAR/10/01	South Africa	2001	[111]	I	AY442912	Not given
SAR/11/01	South Africa	2001	[111]	I	AY442913	Not given
SAR/2/01	South Africa	2001	[111]	I	AY442904	Not given
SAR/3/01	South Africa	2001	[111]	I	AY442905	Not given
SAR/4/01	South Africa	2001	[111]	I	AY442906	Not given
SAR/5/01	South Africa	2001	[111]	I	AY442907	Not given
SAR/6/01	South Africa	2001	[111]	I	AY442908	Not given
SAR/7/01	South Africa	2001	[111]	I	AY442909	Not given
SAR/8/01	South Africa	2001	[111]	I	AY442910	Not given
SAR/9/01	South Africa	2001	[111]	I	AY442911	Not given
SUD/6/77	Sudan	1977	[123]	XIII	AY343939	Cattle
SUD/9/77	Sudan	1977		XIII	AY442014	Cattle
SUD/1/2007	Sudan	2007	[58]	VII	GU566071	Cattle
SUD/1/2008	Sudan	2008	[58]	XIII	GU566072	Cattle
SUD/2/2008	Sudan	2008	[58]	XIII	GU566073	Cattle
SUD/4/2010	Sudan	2010	[59]	VII	KF112968	Cattle
TAN/1/75	Tanzania	1975	[123]	IV	AY343970	Cattle
TAN/1/86	Tanzania	1986	[123]	IV	AY343971	Cattle
TOG/1/90	Togo	1990	[59]	V	KF112970	Cattle
UGA/51/75	Uganda	1975	[123]	XII	AY343963	Cattle

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
UGA/3/76	Uganda	1976	[123]	XII	AY343964	Cattle
UGA/8/76	Uganda	1976	[123]	XII	AY343965	Cattle
U267/83	Uganda	1983	[129]	IV	HM623682	Cattle
UGA/3/91	Uganda	1991	[123]	IX	AY343966	Cattle
UGA/9/95	Uganda	1995	[123]	IX	AY343967	Cattle
UGA/19/98	Uganda	1998	[123]	X	AY343969	Cattle
UGA/28/98	Uganda	1998	[123]	X	AY343968	Cattle
Murchison Falls National Park	Uganda	2002	[24]	VII	FJ461346	*S. caffer*
UGA_002/2002	Uganda	2002		X	JF749862	Not given
UGA/02/02	Uganda	2002		X	DQ009731	Not given
SAT 2 Uga 1/07	Uganda	2007	[4]	X	HM067705	*S. caffer*
SAT 2 Uga 2/07	Uganda	2007	[4]	X	HM067704	*S. caffer*
RHO/1/48	Zambia	1948	[22]	III	AJ251475, AY593847	Not given
ZAM/3/81	Zambia	1981	[59]	III	KF112971	Cattle
ZAM/10/93	Zambia	1993	[9]	III	AF367117	*S. caffer*
ZAM/9/93	Zambia	1993	[9]	III	AF367116	*S. caffer*
ZAM/10/96	Zambia	1996	[9]	III	AF367121	Not given
ZAM/7/96	Zambia	1996	[9]	III	AF367120	Not given
RHO/2/72	Zimbabwe	1972	[59]	I	KF112962	Cattle
ZIM/5/81	Zimbabwe	1981	[54]	II	EF134951	Cattle
ZIM/5/81	Zimbabwe	1981	[59]	II	KF112972	Cattle
ZIM/5/83	Zimbabwe	1983	[108, 149–151]	II	AF540910, DQ009726, JQ639289	Cattle
ZIM/1/87	Zimbabwe	1987	[59]	II	KF112973	Cattle
ZIM/5/87	Zimbabwe	1987	[59]	II	KF112974	Cattle
ZIM/01/88	Zimbabwe	1988	[9, 118]	II	AF367108, GU194491	*S. caffer*
ZIM/2/88	Zimbabwe	1988	[108]	II	JQ639294	*S. caffer*

Continued on next page

Isolate	Country	Year	Reference(s)	Topotype	Accession	Host
ZIM/7/89	Zimbabwe	1989	[108]	II	JQ639296	*S. caffer*
ZIM/8/89	Zimbabwe	1989	[59]	II	KF112975	Cattle
ZIM/9/89	Zimbabwe	1989	[59]	II	KF112976	Cattle
ZIM/34/90	Zimbabwe	1990	[118]	II	GU194490	*S. caffer*
Zim/14/90	Zimbabwe	1990		II	DQ009728	Not given
ZIM/Gn10/91	Zimbabwe	1991	[9]	I	AF367113	*S. caffer*
ZIM/10/91	Zimbabwe	1991	[118]	I	GU194493	*S. caffer*
ZIM/17/91	Zimbabwe	1991		II	DQ009727	*S. caffer*
ZIM/8/94	Zimbabwe	1994	[108, 118]	I	GU194492, JQ639290	*S. caffer*
ZIM/4/97	Zimbabwe	1997	[108]	II	JQ639293	Cattle
ZIM/44/97	Zimbabwe	1997	[108]	II	JQ639291	*S. caffer*
ZIM/267/98	Zimbabwe	1998	[9]	II	AF367130	*S. caffer*
ZIM/1/00	Zimbabwe	2000	[9]	II	AF367136	*S. caffer*
ZIM/13/01	Zimbabwe	2001	[108]	II	JQ639292	Cattle
ZIM/5/02	Zimbabwe	2002	[108]	II	JQ639295	Cattle
ZIM_22/2003	Zimbabwe	2003		I	JF749864	Not given

Table A.1: All FMDV serotype SAT 2 isolates used in chapter 2. Accession numbers are for the NCBI Nucleotide database.

Isolate	Country	Date of collection	Location of collection within country	Topotype	Accession	Host
BOT/3/77	Botswana	September 1977	Not recorded	III	KF112928	Cattle
BOT 08/78	Botswana	August 1978	Not recorded	III	KF112929	Cattle
BUN/4/86	Burundi	1986	Not recorded	VIII	KF112930	Not given
EGY/16/2012	Egypt	8 March 2012	Abo Greda, Farasqour, Domyat, Delta	VII	KF112931	*B. bubalis*
EGY/17/2012	Egypt	8 March 2012	Abo Greda, Farasqour, Domyat, Delta	VII	KF112932	*B. bubalis*
EGY/21/2012	Egypt	13 March 2012	Fwa, Sendion, Kafr El Sheikh, Delta	VII	KF112933	Cattle
EGY/22/2012	Egypt	13 March 2012	Fwa, Sendion, Kafr El Sheikh, Delta	VII	KF112934	Cattle
EGY/28/2012	Egypt	17 May 2012	Manzala, El Daqahleya, Delta	VII	KF112935	Cattle
EGY/29/2012	Egypt	19 May 2012	Rawda, Faras Qour, Domyat, Delta	VII	KF112936	Cattle
EGY/31/2012	Egypt	21 May 2012	Kafr El Tagi, Kafr El Sheikh, Kafr El Sheikh, Delta	VII	KF112937	Cattle
ETH/42/2009	Ethiopia	16 June 2009	Lare, Nuer Zone, Gambela Region	XIII	KF112938	Not given
ETH/48/2009	Ethiopia	12 August 2009	Gambela, Anuak Zone. Gambela Region	XIII	KF112939	Not given
ETH/51/2009	Ethiopia	5 November 2009	Kinbibit, North Shewa, Oromia Region	XIII	KF112940	Cattle
ETH/52/2009	Ethiopia	17 November 2009	Kinbibit, North Shewa, Oromia Region	XIII	KF112941	Cattle
ETH/53/2009	Ethiopia	17 November 2009	Kinbibit, North Shewa, Oromia Region	XIII	KF112942	Cattle
ETH/56/2009	Ethiopia	2 December 2009	Mulo, Oromia Region	XIII	KF112943	Cattle
ETH/64/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112944	Cattle
ETH/65/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112945	Cattle
ETH/67/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112946	Cattle
ETH/68/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112947	Cattle
ETH/69/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112948	Cattle
ETH/70/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112949	Cattle
ETH/72/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112950	Pig
ETH/73/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112951	Cattle
ETH/74/2009	Ethiopia	19 November 2009	Debre Berhan, North Shewa, Amhara Region	XIII	KF112952	Cattle
ETH/75/2009	Ethiopia	29 November 2009	Sululta, Oromia Region	XIII	KF112953	Cattle

Continued on next page

Isolate	Country	Date of collection	Location of collection within country	Topotype	Accession	Host
ETH/76/2009	Ethiopia	29 November 2009	Sululta, Oromia Region	XIII	KF112954	Cattle
ETH/77/2009	Ethiopia	29 November 20099	Sululta, Oromia Region	XIII	KF112955	Cattle
ETH/2/2010	Ethiopia	25 January 2010	Debre Zeit, East Shewa, Oromia Region	XIII	KF112956	Cattle
IVY/3/90	Côte d'Ivoire	1990	Bingerville	V	KF112957	Cattle
KNP/1/85	South Africa	21 November 1985	Gudzane, Kruger National Park	I	KF112958	*A. melampus*
MOZ/1/70	Mozambique	10 January 1970	Manica	I	KF112959	Not given
NGR/15/2005	Niger	2005	Not recorded	VII	KF112960	Cattle
NYE/29/90	North Yemen	July 1990	Sana'a abattoir	IV	KF112961	Cattle
RHO/2/72	Zimbabwe	25 September 1972	Zaka TTL	I	KF112962	Cattle
RWA/13/96	Rwanda	6 May 1996	Runda	VIII	KF112964	Not given
RWA/1/2004	Rwanda	2004	Not recorded	VIII	KF112963	Cattle
SA/2/67	South Africa	9 October 1967	Kruger National Park	I	KF112965	*A. melampus*
SAR/3/79	South Africa	22 June 1979	Maswanganye Giuani	I	KF112966	Cattle
SEN/27/2009	Senegal	9 October 2009	Ross Bethio, St. Louis	VII	KF112967	Cattle
SWA/4/89	Namibia	12 November 1989	Sigwe village, East Caprivi	III	KF112969	Cattle
SUD/4/2010	Sudan	9 February 2010	Sheikan, Sheikan, North Kordafan	VII	KF112968	Cattle
TOG/1/90	Togo	November 1990	Not recorded	V	KF112970	Cattle
ZAM/3/81	Zambia	4 November 1981	Kambwa village, Monze district	III	KF112971	Cattle
ZIM/5/81	Zimbabwe	6 November 1981	Lubu Diptank, Manjolo TTL	II	KF112972	Cattle
ZIM/1/87	Zimbabwe	24 March 1987	Insiza, Matabeleland North	II	KF112973	Cattle
ZIM/5/87	Zimbabwe	1 July 1987	Triangle, Chiredzi, Masvingo	II	KF112974	Cattle
ZIM/8/89	Zimbabwe	27 April 1989	Mutorashanga, Mashonaland West	II	KF112975	Cattle
ZIM/9/89	Zimbabwe	4 May 1989	Gweru, Midlands	II	KF112976	Cattle

Table A.2: Further information about the 49 newly-sequenced SAT 2 isolates used in chapter 2. Accession numbers are for the NCBI Nucleotide database.

Appendix B

Full results of analysis of all sampling replicates of foot-and-mouth disease serotype O sequences

B.1 Introduction

This appendix presents in fuller detail the results of all the sampling replicates conducted in the analysis of foot-and-mouth disease virus (FMDV) serotype O presented in chapter 4. In addition, it presents the results of a separate analysis of the host species datasets for topotypes ME-SA and SEA conducted using the BASTA approximate structured coalescent method recently developed by De Maio et al. [30], and compares those results with those obtained from the continuous-time Markov chain method used in the chapter. While the CTMC model assumes that all lineages are part of a single, freely-mixing population and treats host as

a characteristic of each isolate that "mutates" independently of that population structure, BASTA assumes that the population of lineages infecting each type of host is a separate "deme" (of constant size) and that lineages migrate from one deme to another at fixed rates.

B.2 Results

B.2.1 Analysis of the full serotype

The ten replicates of the sampling scheme for the random local molecular clock analysis of a random sample of 233 serotype O sequences are designated ALL1 to ALL10. ALL4 is the replicate presented in the main text. Figure B.1 displays the posterior distributions for the time to most recent common ancestor (TMRCA) of the full serotype in each case, and figure B.2 the reconstructed skygrid plots. Figure B.3 is every MCC tree, with branches coloured by posterior median nucleotide substitution rate. Notably, replicate ALL10 has a much earlier estimated TMRCA, in December 1865 (March 1851-May 1881), and also uniquely suggests the existence of rate change points separating SEA and the African topotypes from the rest of the tree. The posterior median clock rate for SEA tips in that replicate is around 5×10^{-3} substitutions per site per year, and those for all four African topotypes is around 3.8×10^{-3}. On the other side of the breakpoint, a rate of around 2.5×10^{-3} continues to the root of the tree and is also that on the Euro-SA(2) and (3) tips; there is another change point in the Euro-SA(1) clade and those sequences that are not very closely related to O_1 Campos/58 have tip rates of about 3.2×10^{-3}. Rates in the deep branches are rather slower than in other replicates, which is consistent with the earlier TMRCA estimate.

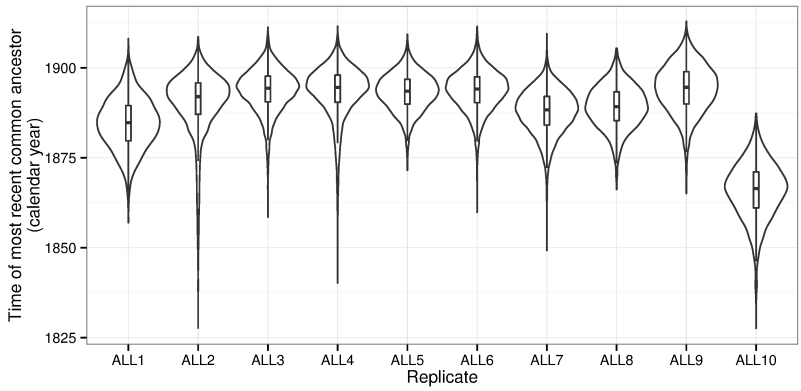


Figure B.1: Violin plots for the posterior distribution of the TMRCA of all type O sequences in each sampling replicate of the analysis of the full serotype.

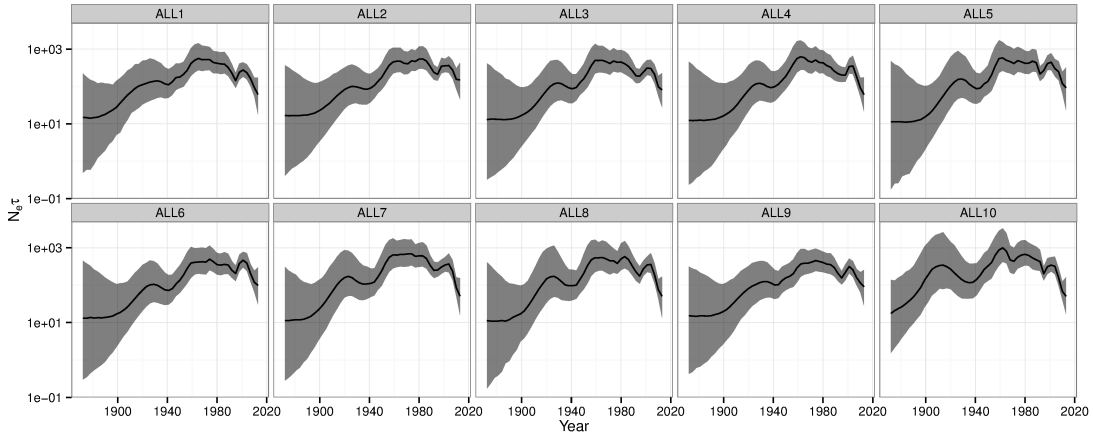
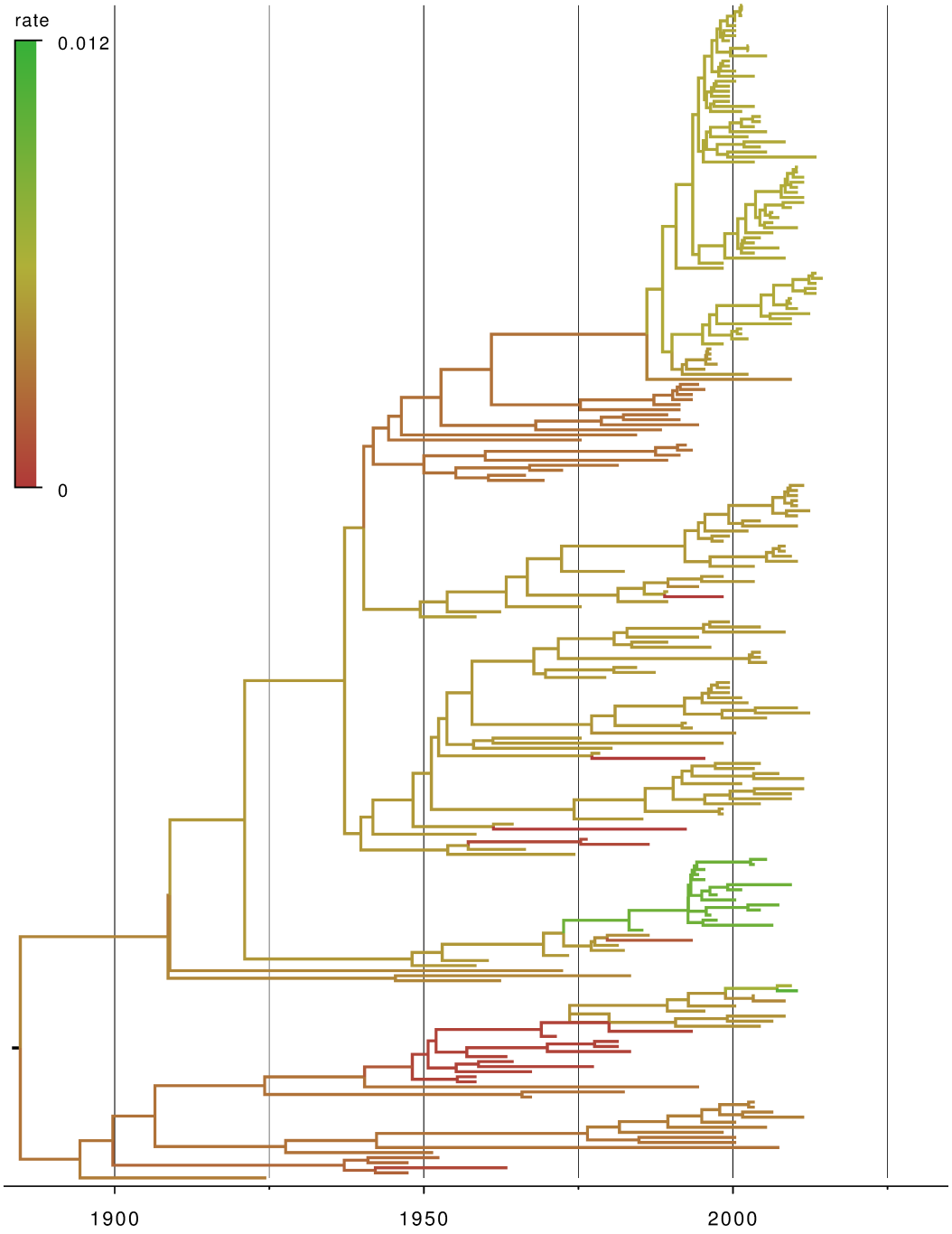
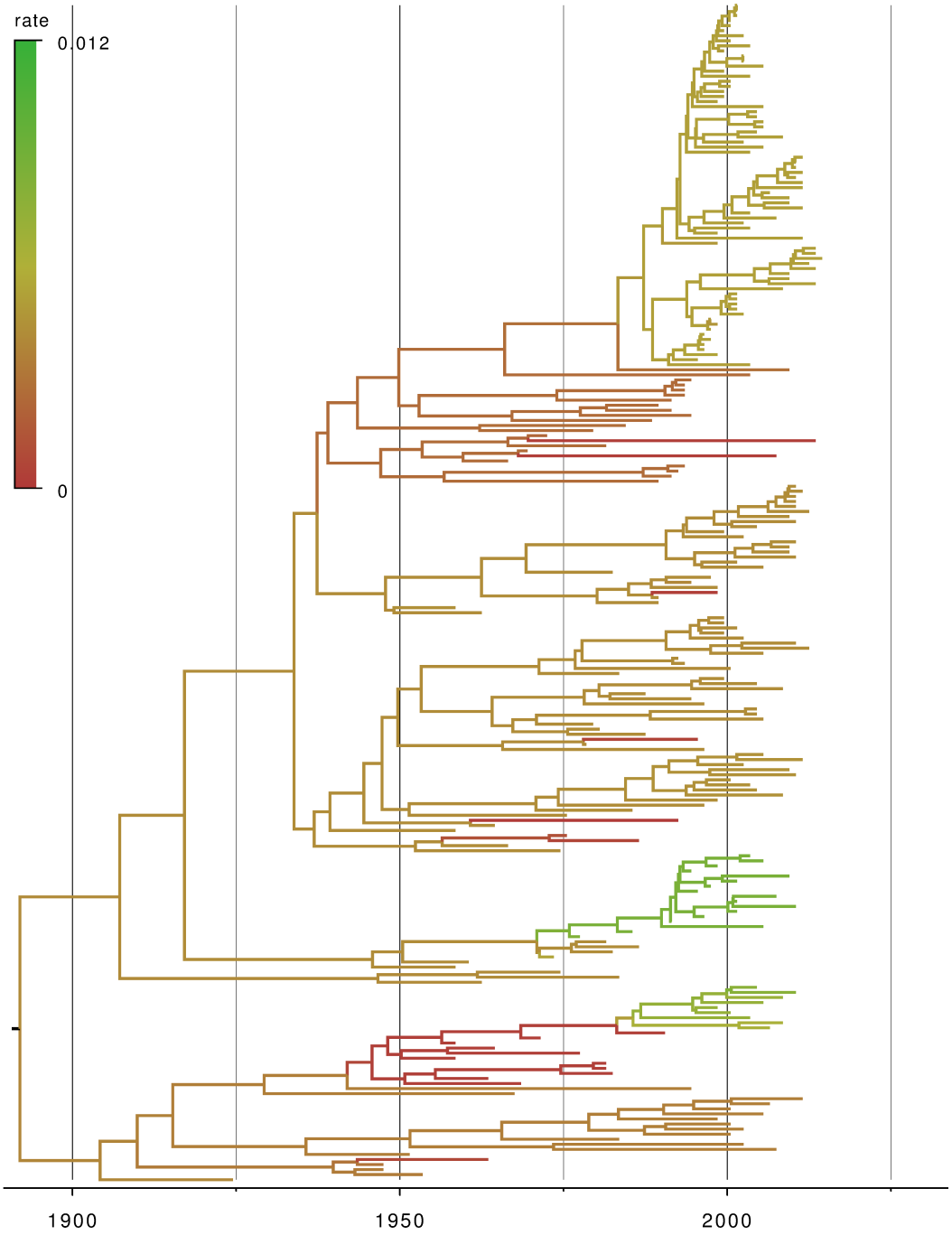


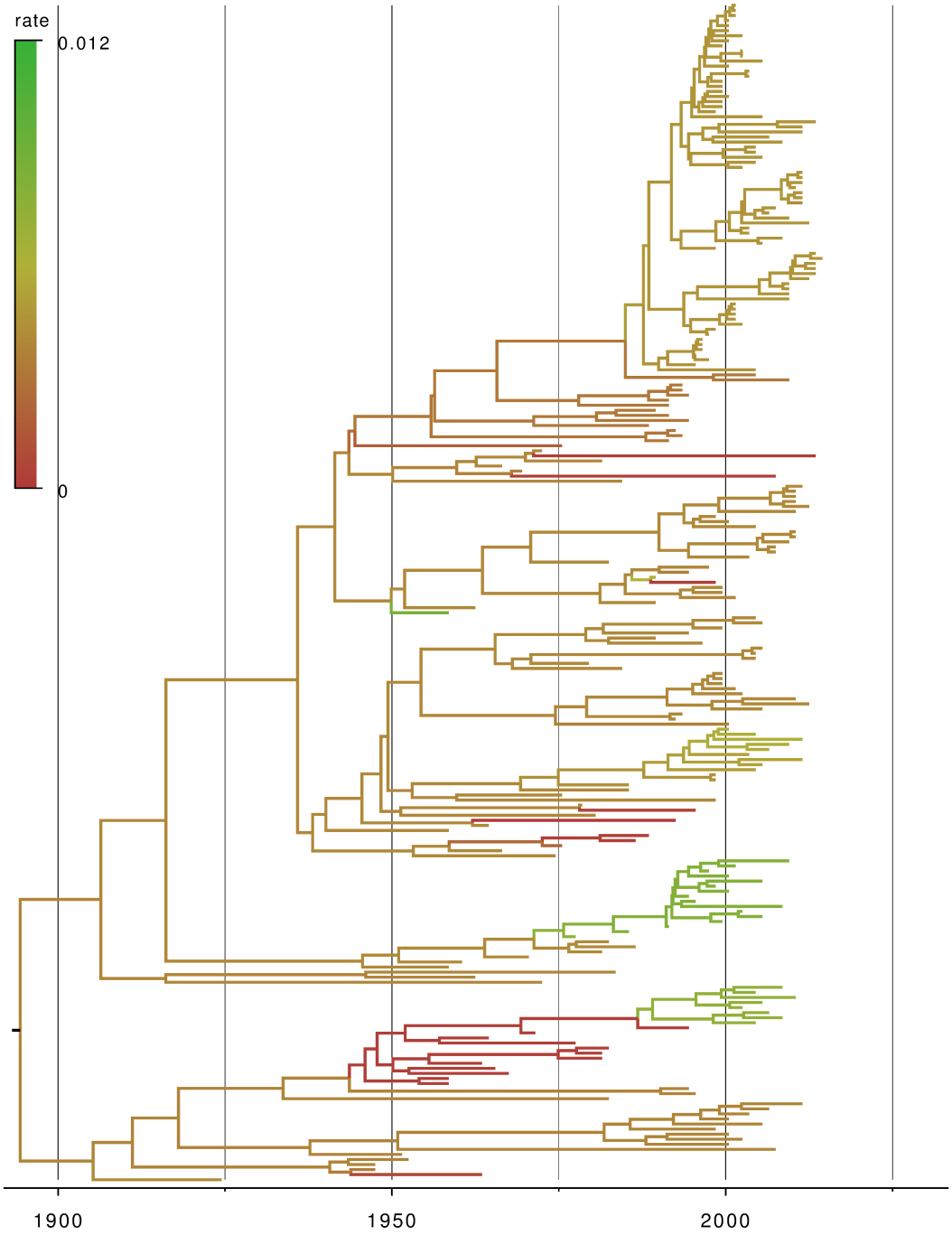
Figure B.2: Reconstructed skygrid plots for each sampling replicate of the analysis of the full serotype. The black line is the median effective population size and the grey area the 95% highest posterior density region.



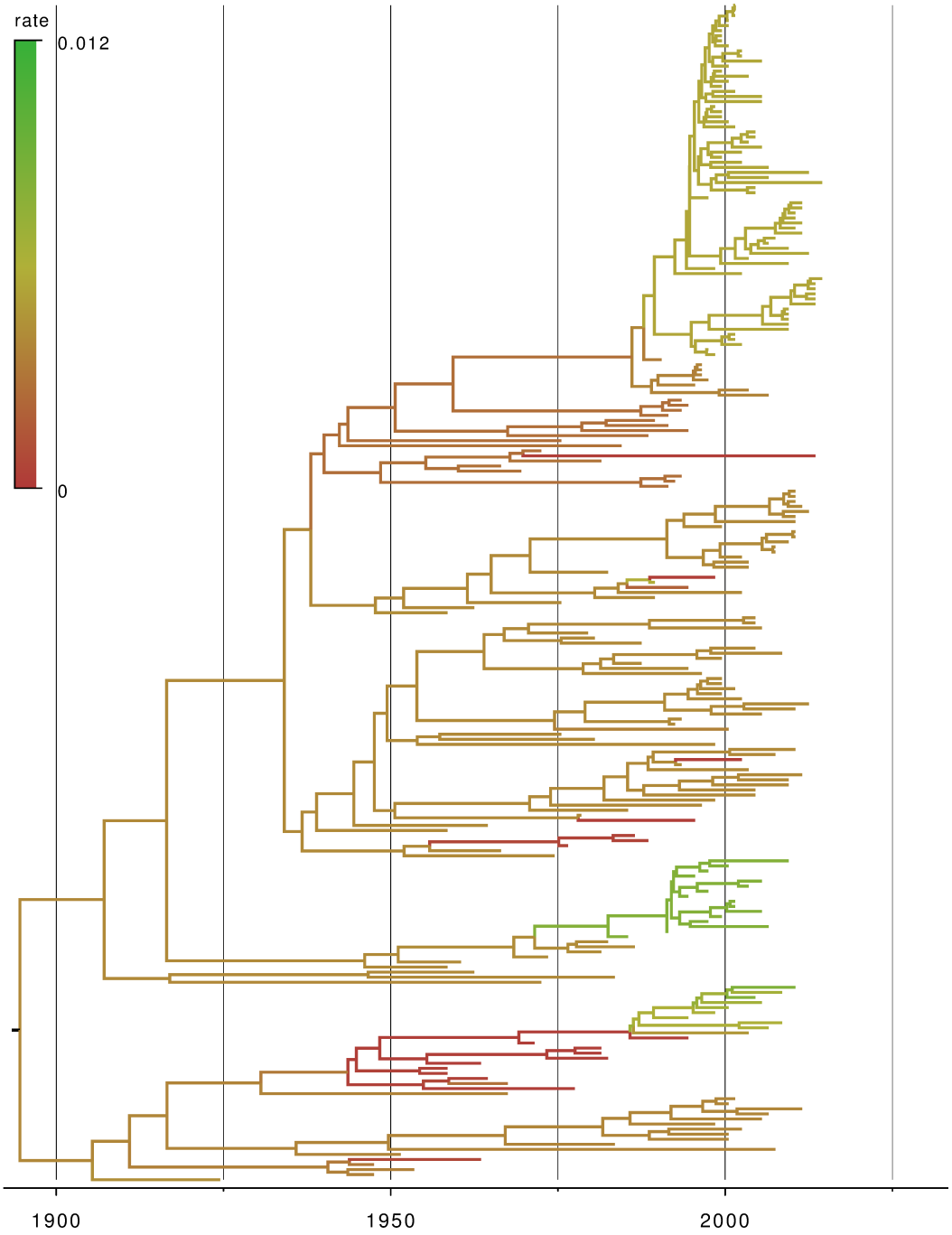
(a) ALL1



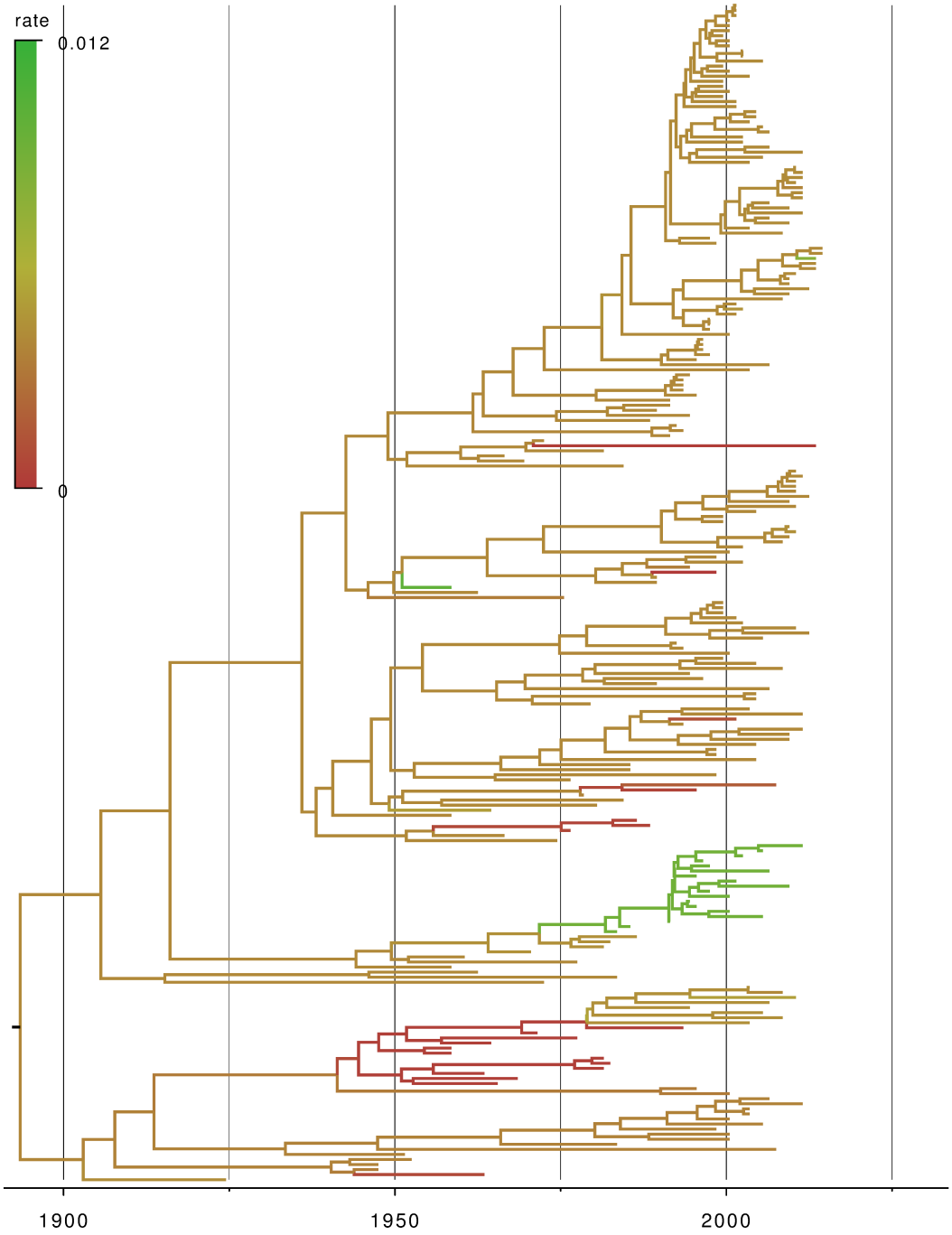
(b) ALL2



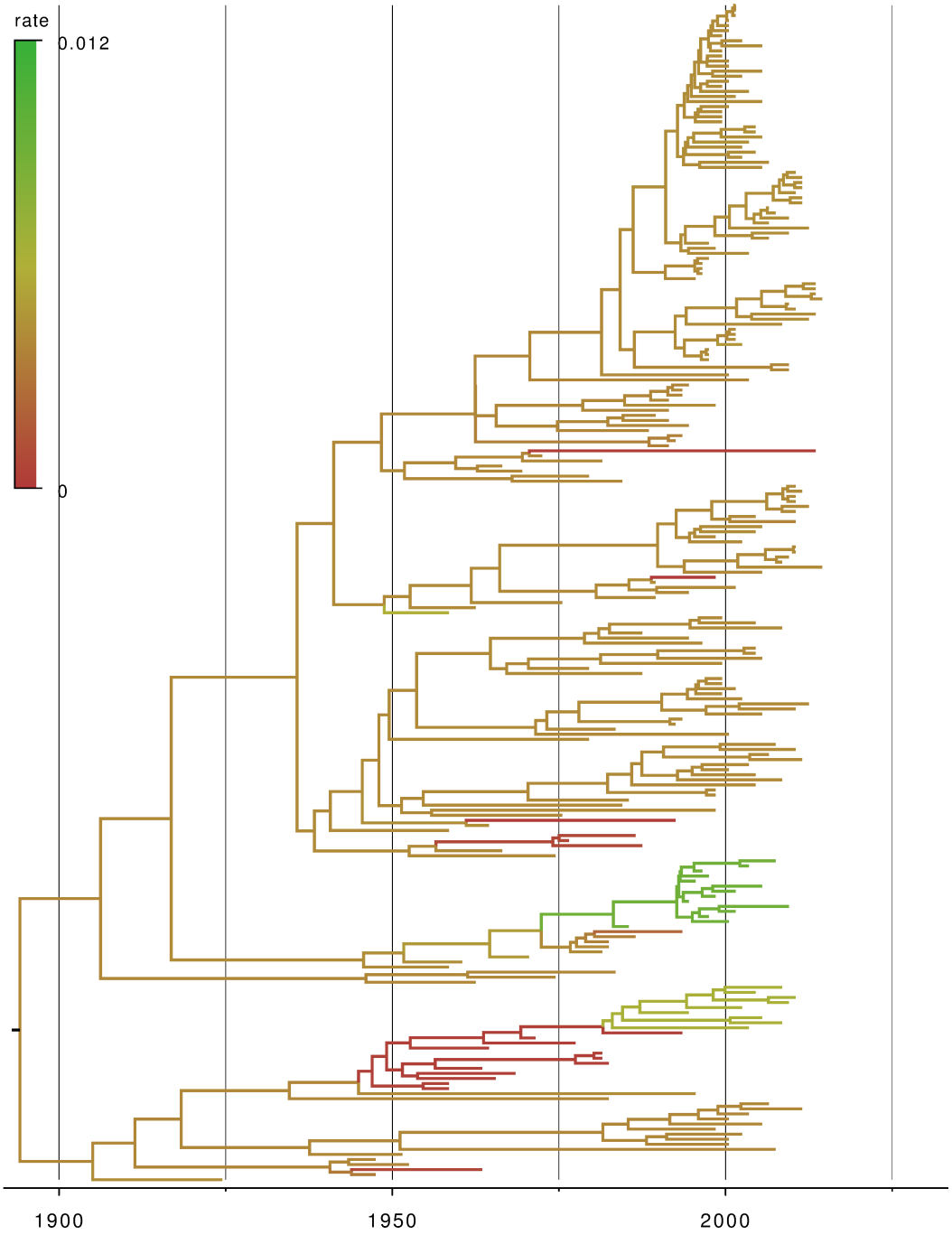
(c) ALL3



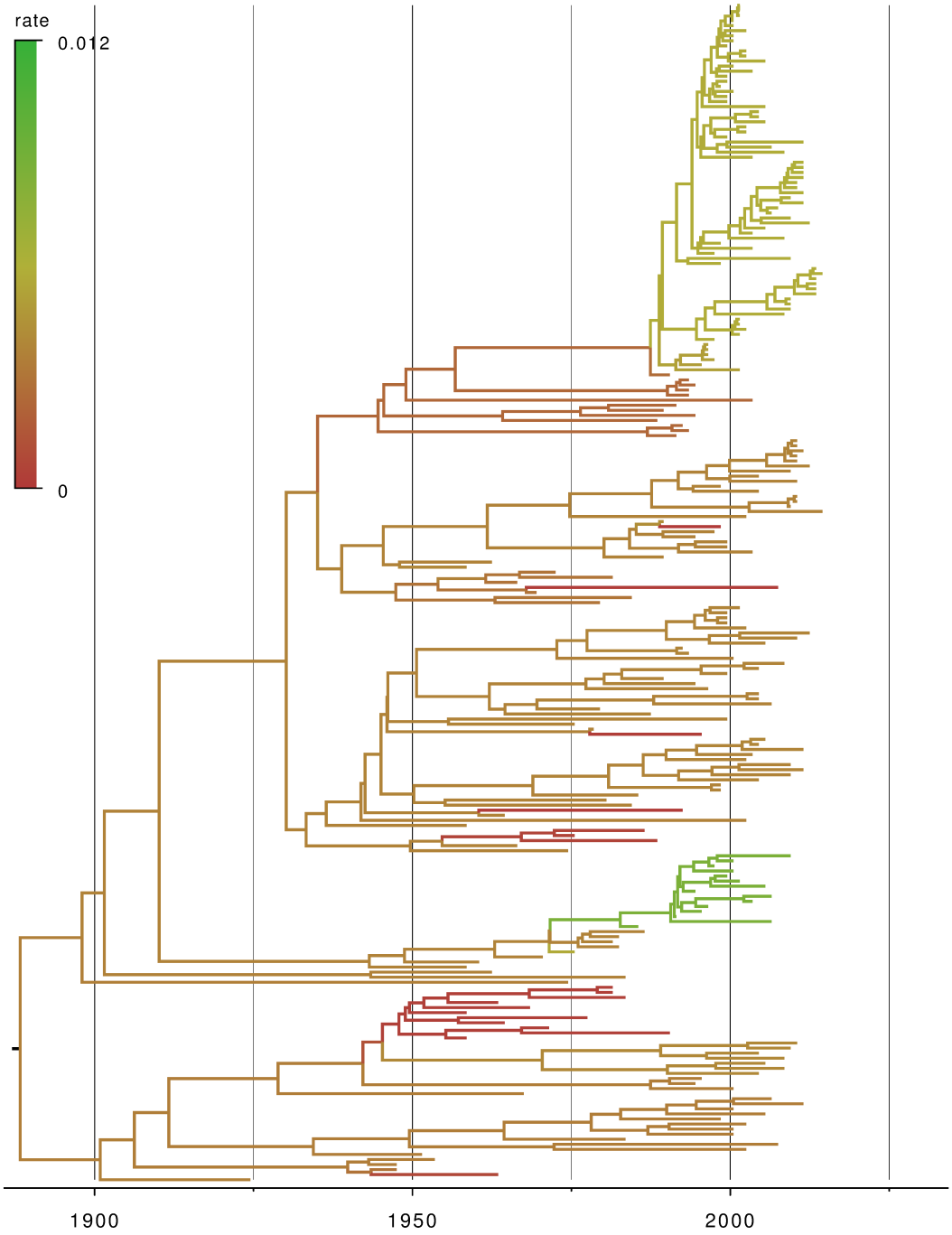
(d) ALL4



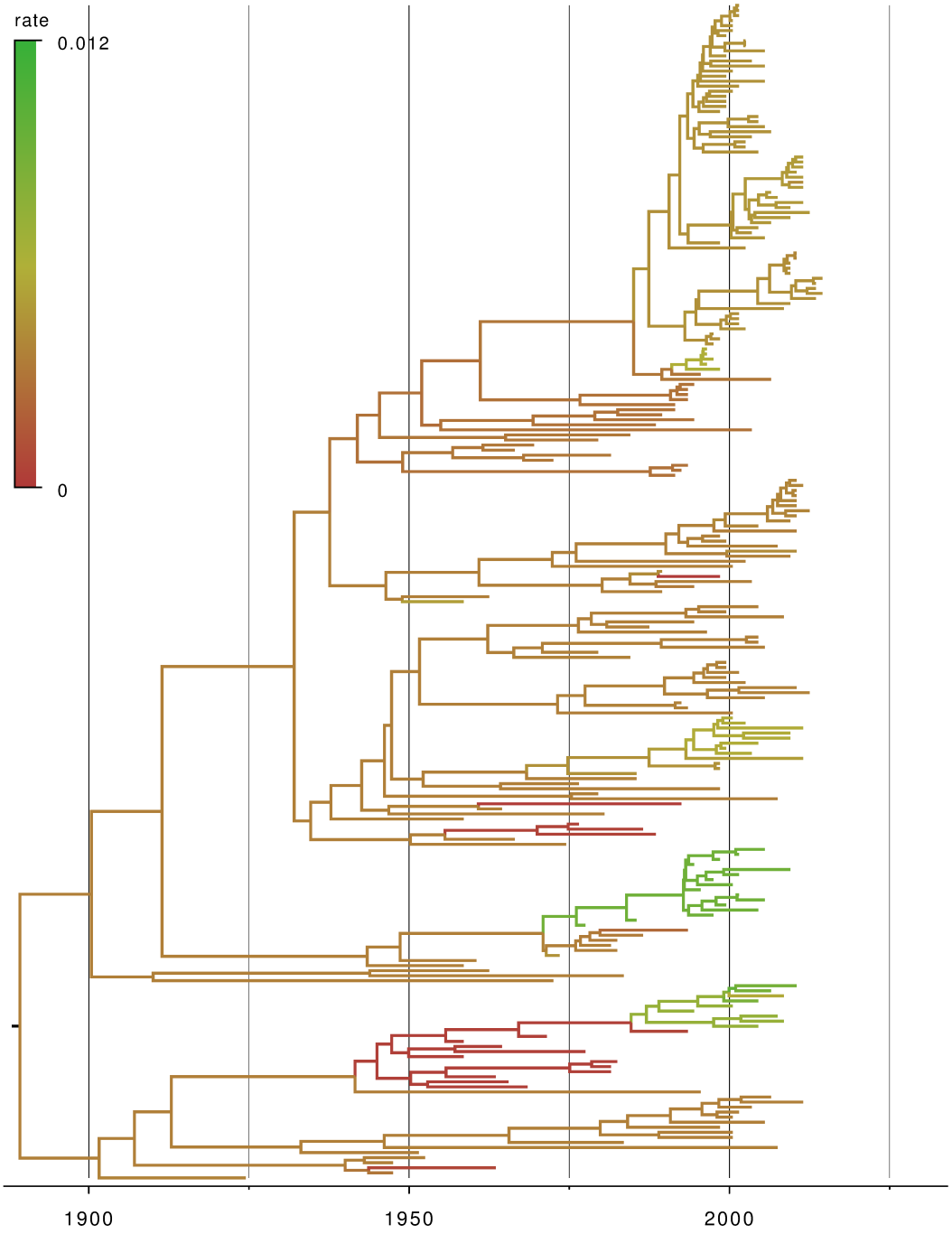
(e) ALL5



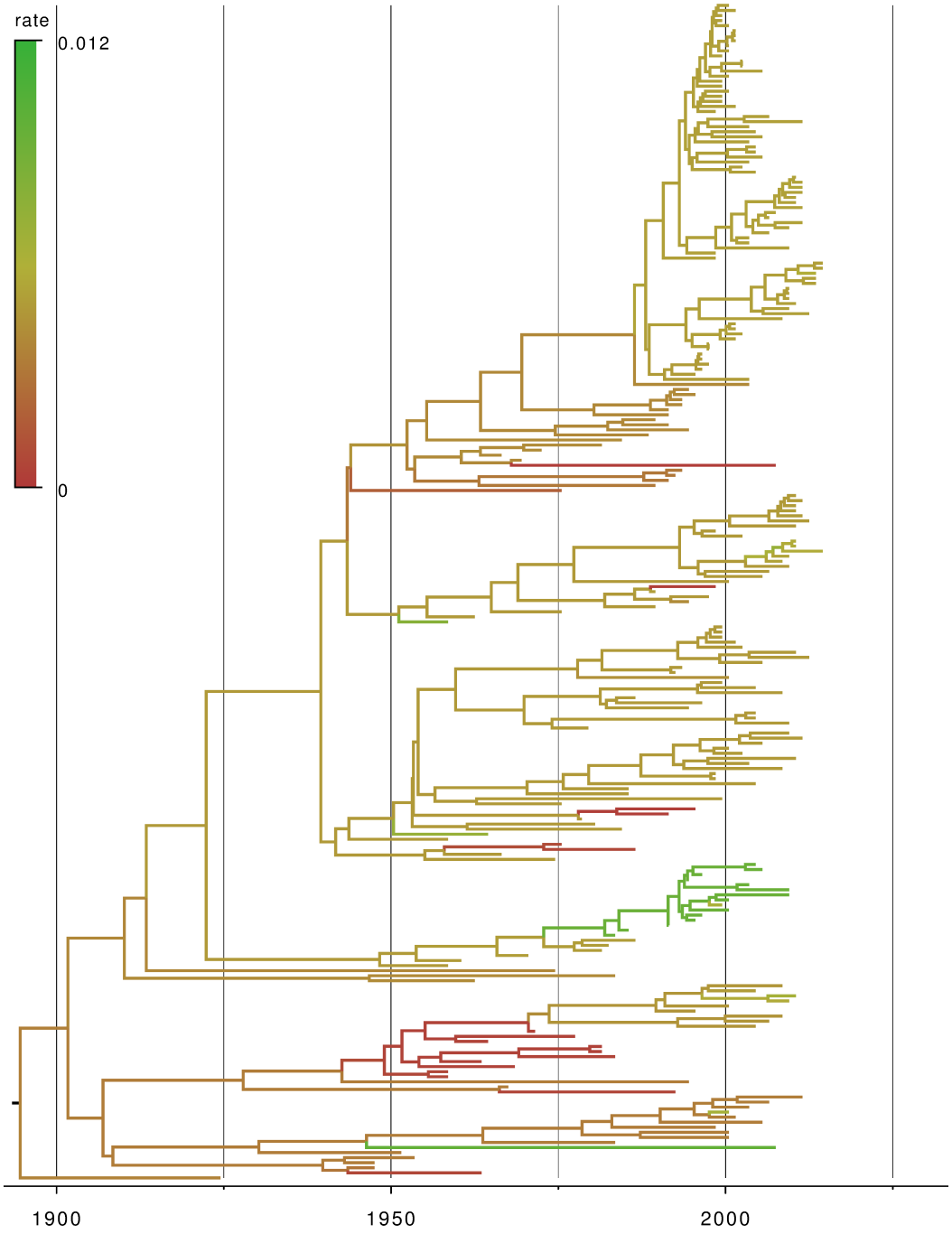
(f) ALL6



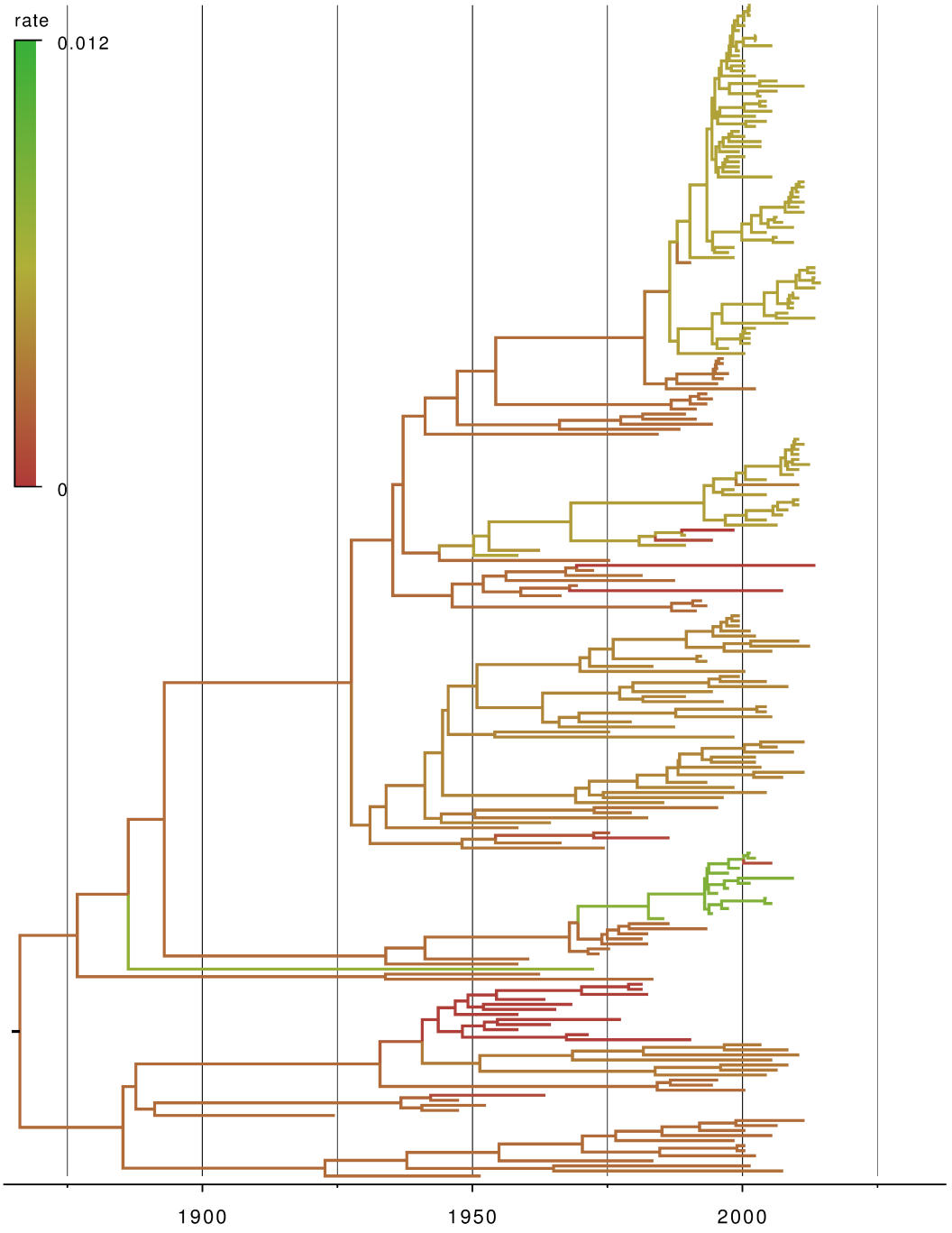
(g) ALL7



(h) ALL8



(i) ALL9



(j) ALL10

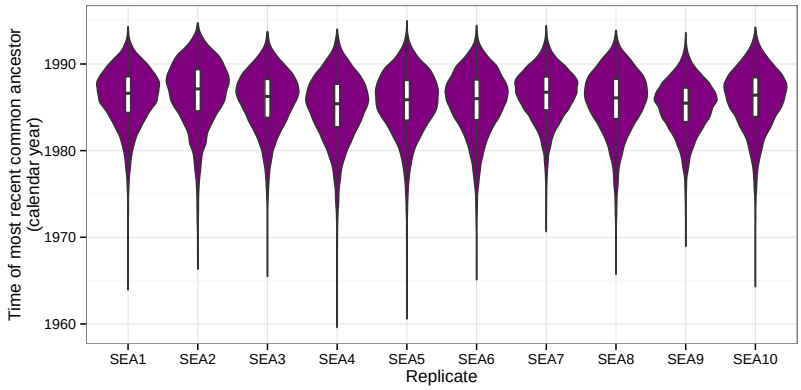
Figure B.3: Maximum clade credibility trees for each sampling replicate of the analysis of the full serotype. Branches are coloured by posterior median molecular clock rate.

B.2.2 Topotype SEA

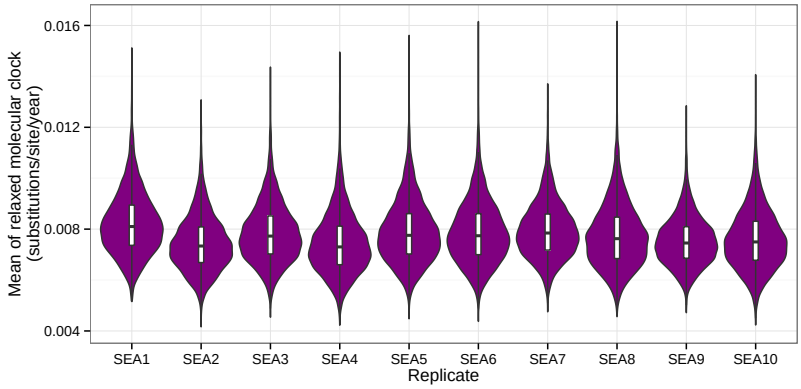
The ten replicates of the phylogeography sampling scheme are designated SEA1 to SEA10 and the ten of the host species sampling scheme SEAH1 to SEAH10. SEA1 and SEAH1 are presented in chapter 4. Figure B.4 depicts the posterior distributions for the TMRCA of all the included sequences and the parameters of the molecular clock. Figure B.5 displays the reconstructed skygrid plots, figure B.6 the posterior distributions for the geographical location of the root of the tree, and figure B.7 the MCC trees and GLM predictor results. There is very little in the way of variation between any of these outputs.

The estimated posterior distributions of the effective population sizes of the demes comprising viruses infecting cattle, pigs and *B. bubalis* from the BASTA analysis are summarised in figure B.8; note that tick increments on the y-axis occur on a log scale. These results are somewhat complex, and for clarity the results are also given in table B.1. The cattle deme always has the largest size in terms of both summary statistics, but the posterior distribution for its size is extremely skewed, as can be seen by the enormous upper limits to the HPD intervals; in some cases the skew is so severe that the posterior mean is actually outside the HPD. The *B. bubalis* deme has a higher posterior median size than the pig deme in all replicates except one (SEAH7).

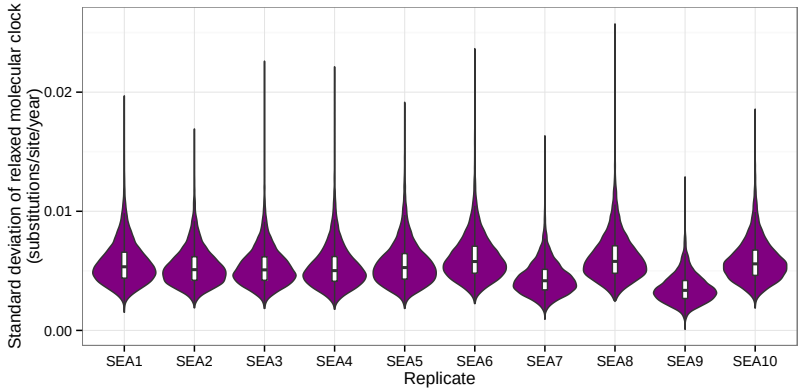
Figure B.9 shows the posterior distributions of the host assigned to the root of the phylogeny, from both analyses. These differ greatly between the two, with the CTMC model preferring cattle and BASTA buffalo or, for one replicate, pigs. Posterior median TMRCAs of all sequences were also always earlier for BASTA (figure B.10), while estimates from CTMC analyses were very similar to those from the phylogeography analysis, those from BASTA ranged from July 1978 (SEAH3, November 1963-March 1988) to October 1982 (SEAH1, August 1971-July 1990). The reason for this does not appear to be a difference in the mean molecular clock



(a) TMRCA



(b) Molecular clock mean



(c) Molecular clock standard deviation

Figure B.4: Violin plots for the posterior distribution of a) the TMRCA of all sequences, b) the mean and c) the standard deviation of the uncorrelated lognormal molecular clock in each sampling replicate of the analysis of the SEA topotype.

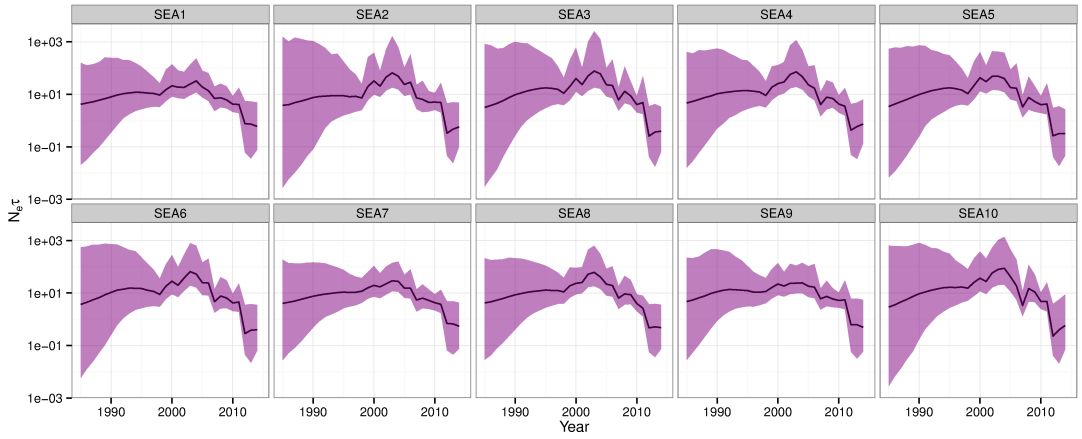


Figure B.5: Reconstructed skygrid plots for each sampling replicate of the analysis of the SEA topotype. The black line is the median effective population size and the grey area the 95% highest posterior density region.

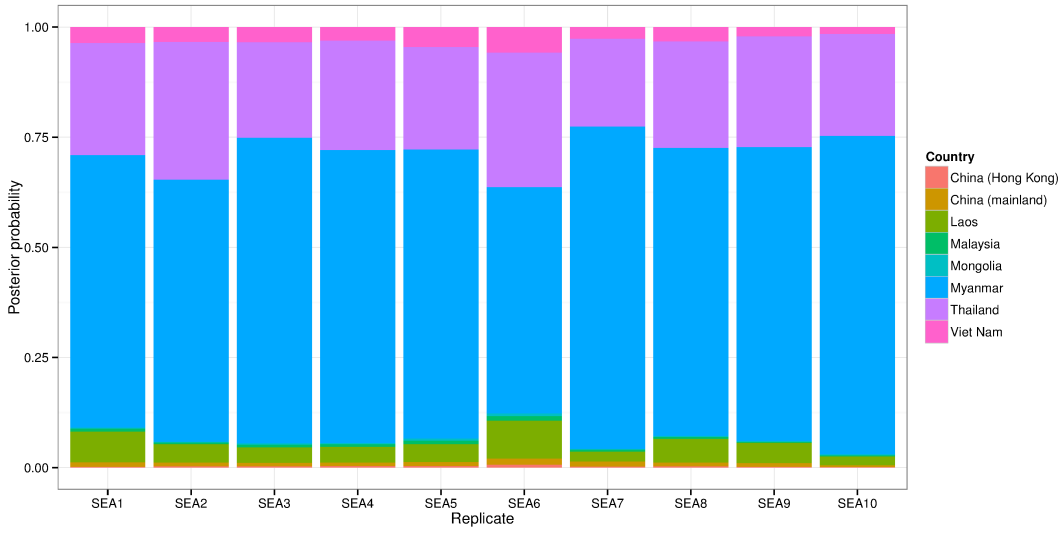
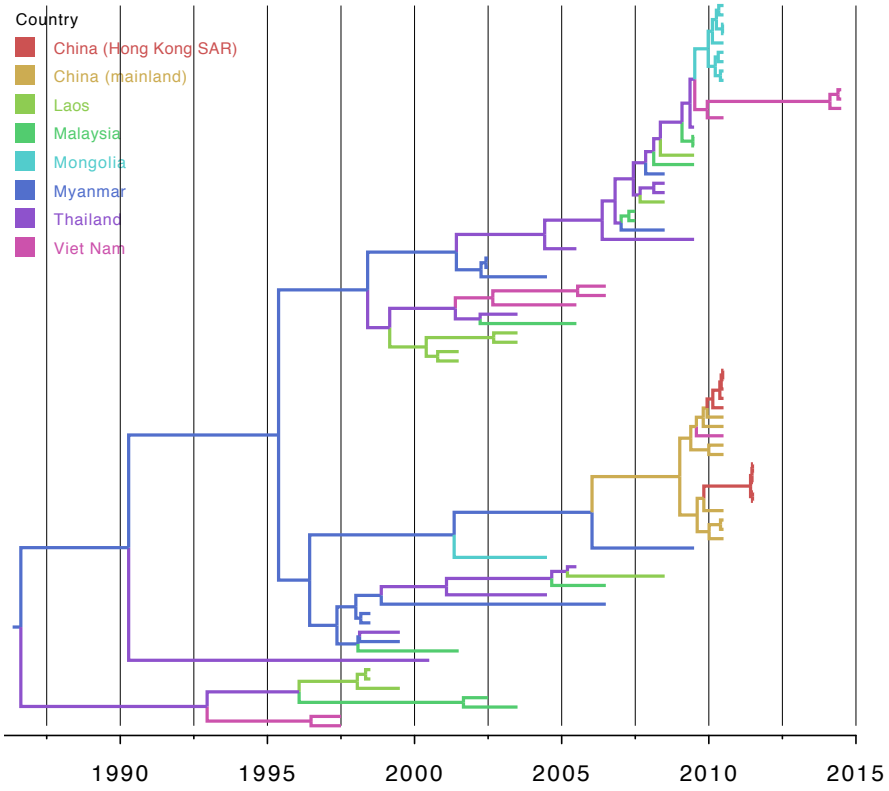
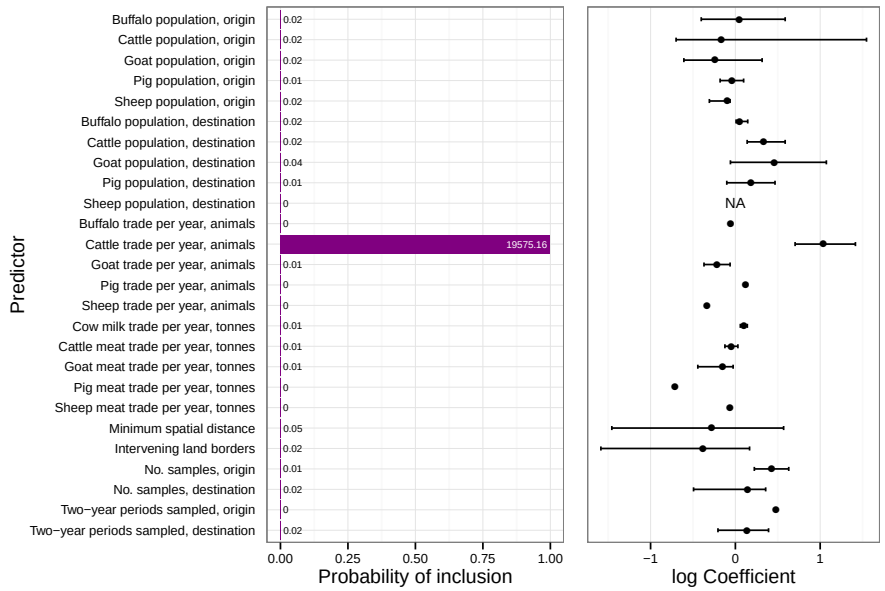


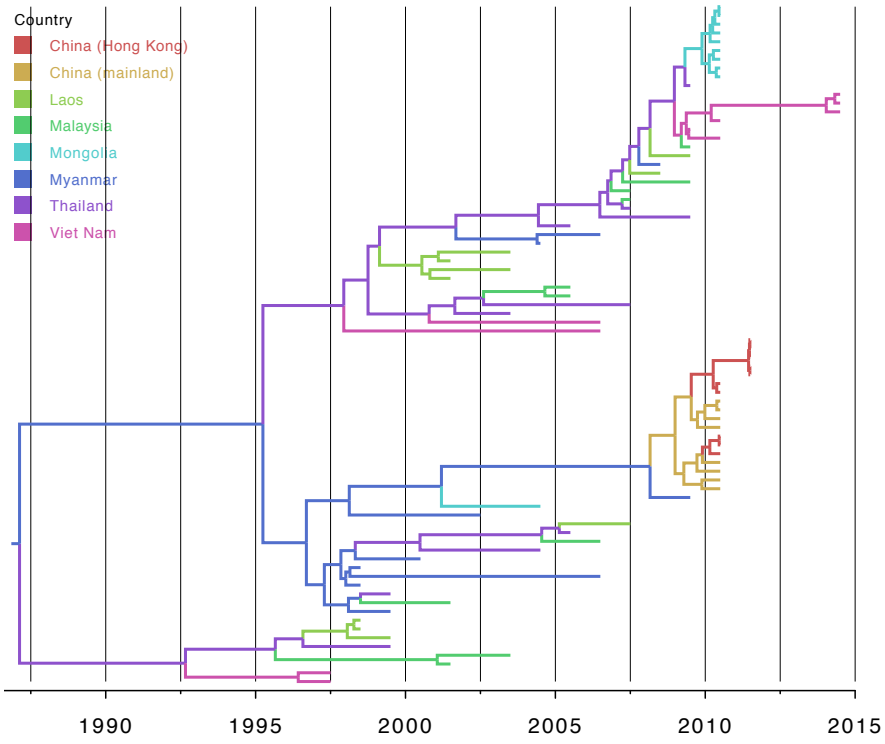
Figure B.6: Posterior distributions for the geographical location of the root node for each sampling replicate of the analysis of the SEA topotype.



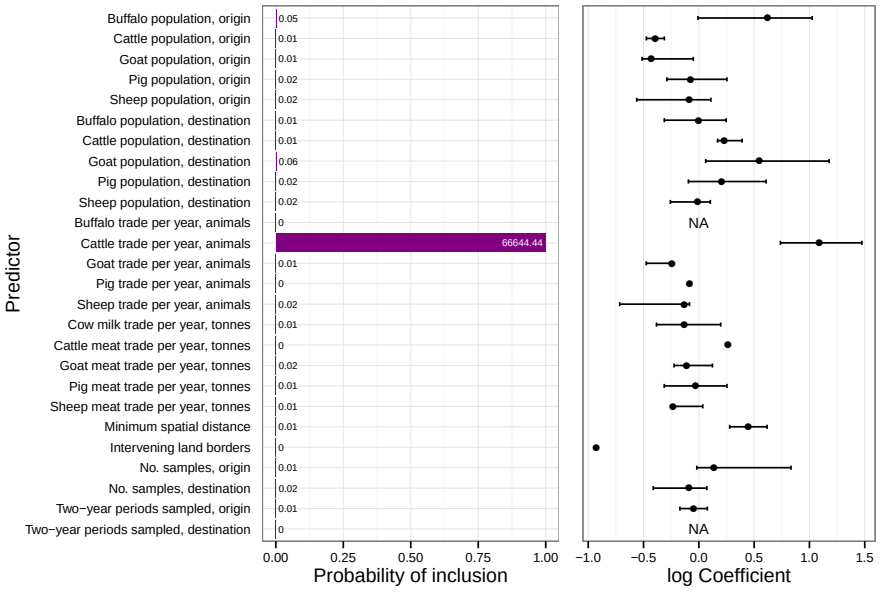
(a) MCC tree, SEA1



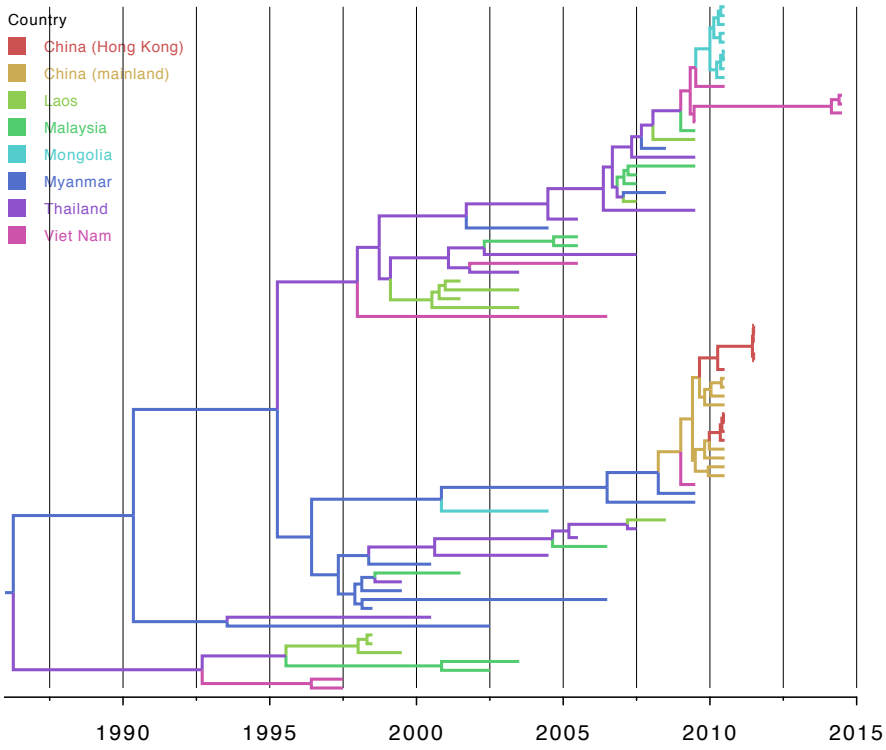
(b) GLM results, SEA1



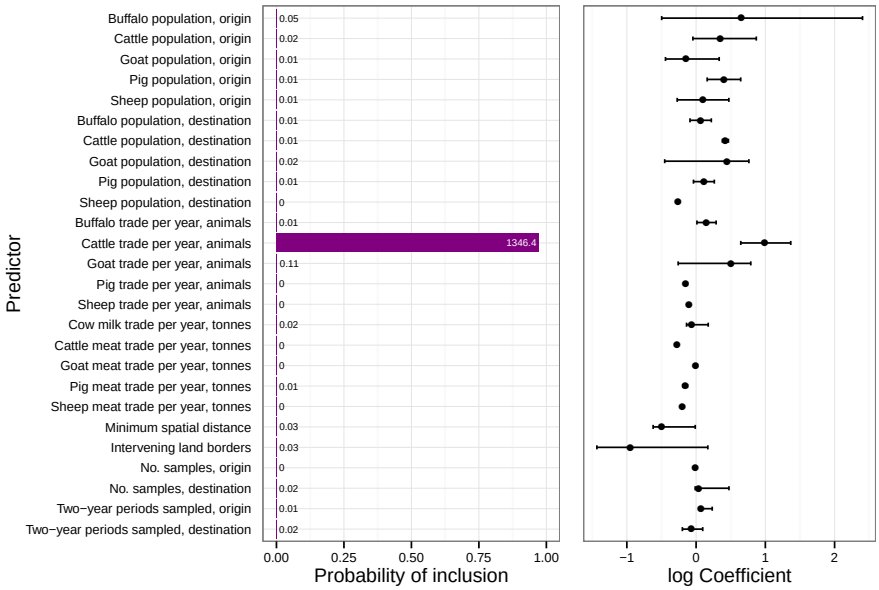
(c) MCC tree, SEA2



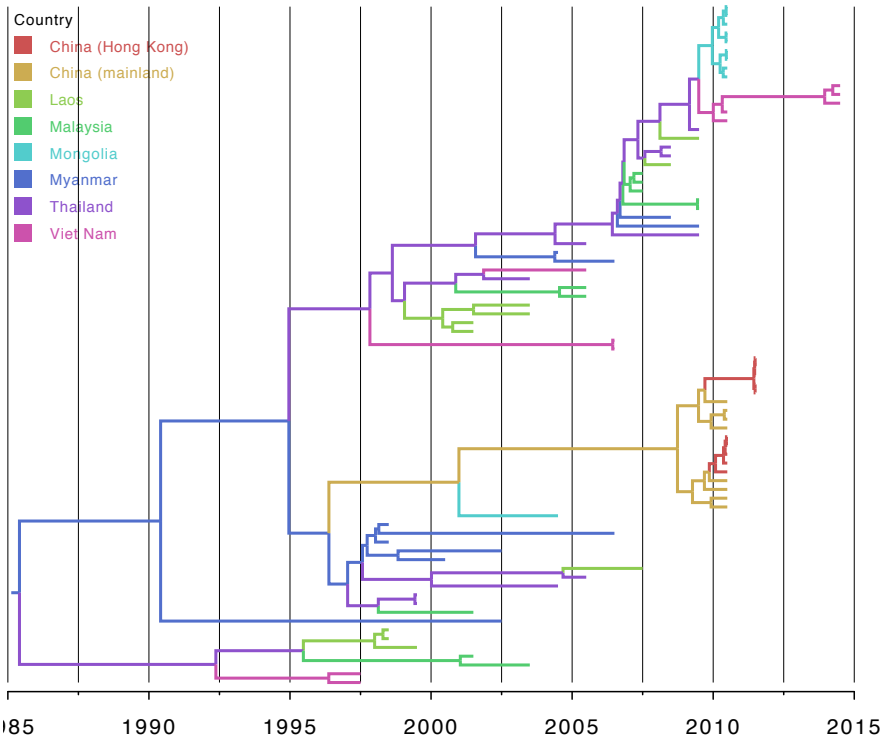
(d) GLM results, SEA2



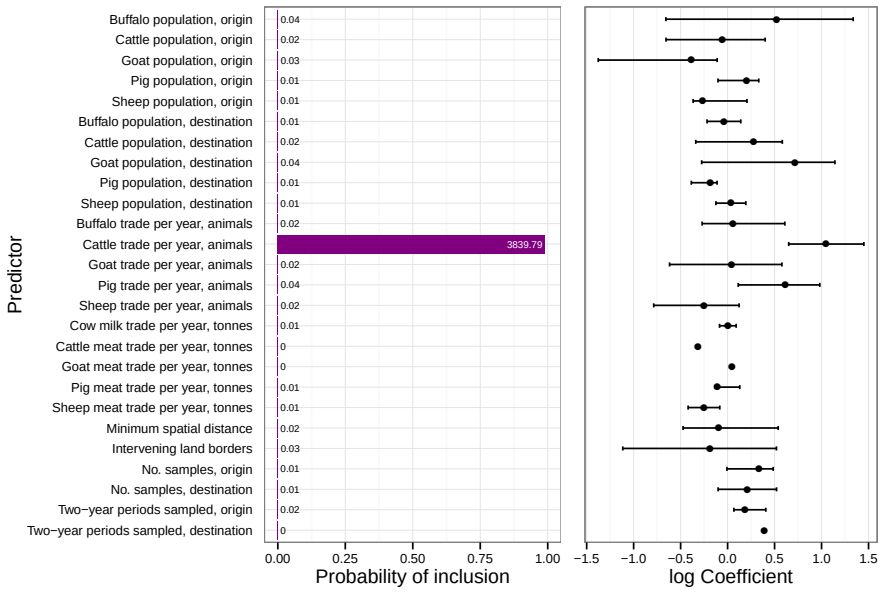
(e) MCC tree, SEA3



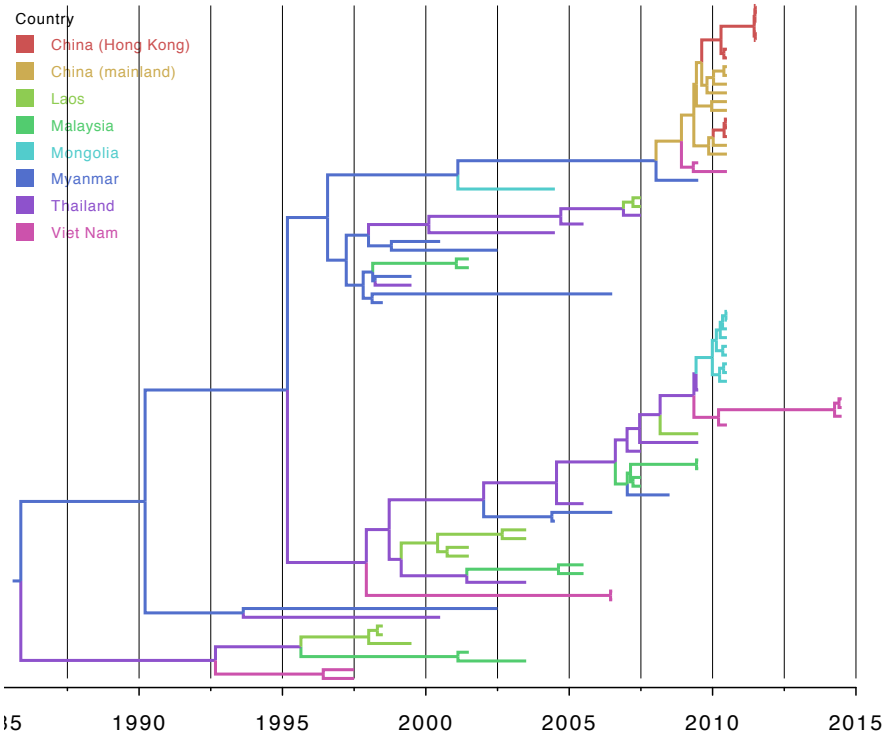
(f) GLM results, SEA3



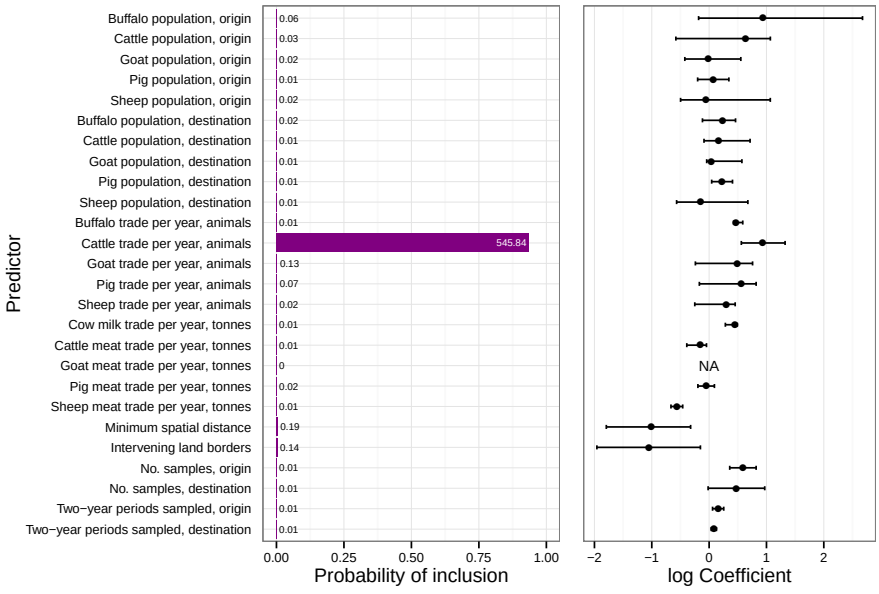
(g) MCC tree, SEA4



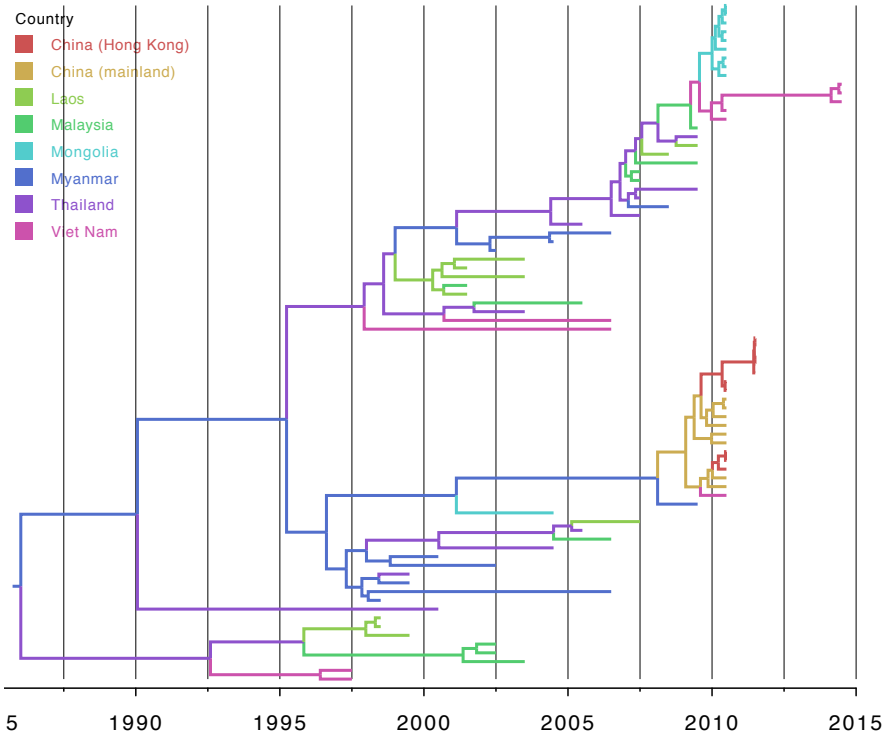
(h) GLM results, SEA4



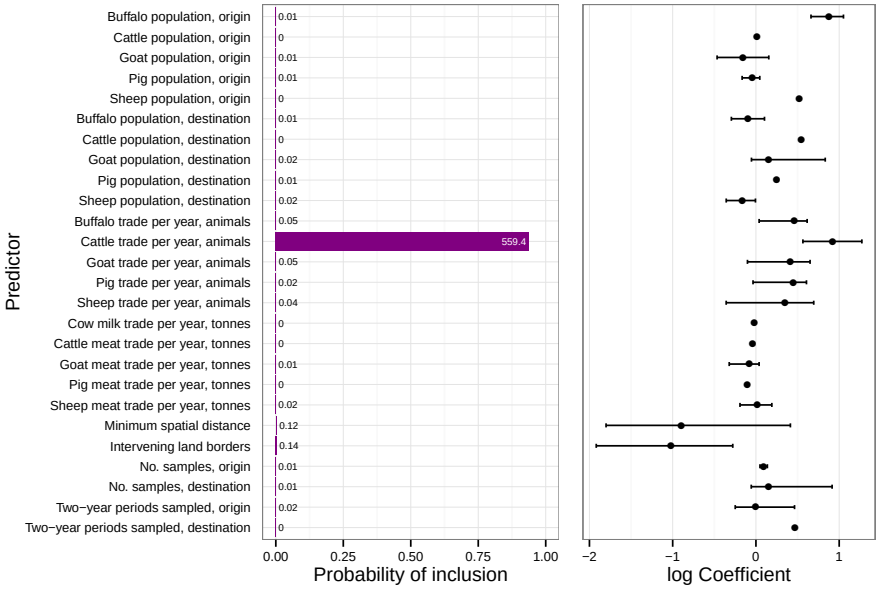
(i) MCC tree, SEA5



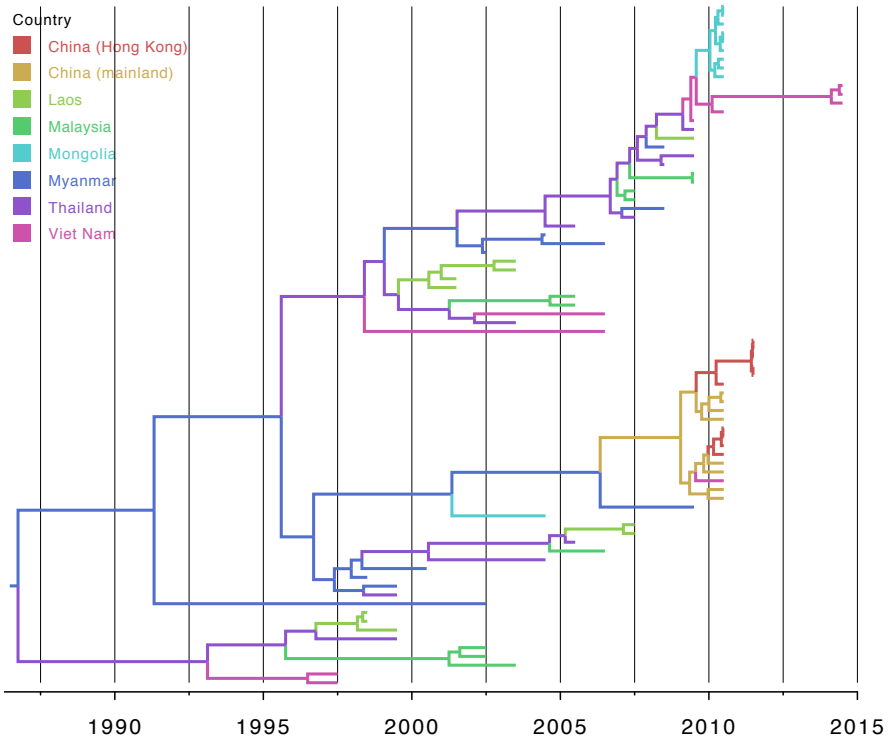
(j) GLM results, SEA5



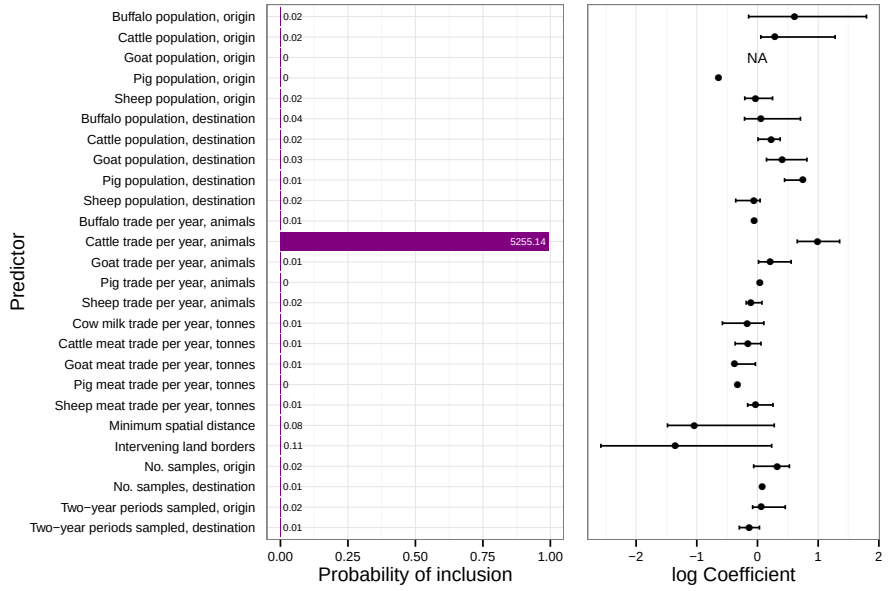
(k) MCC tree, SEA6



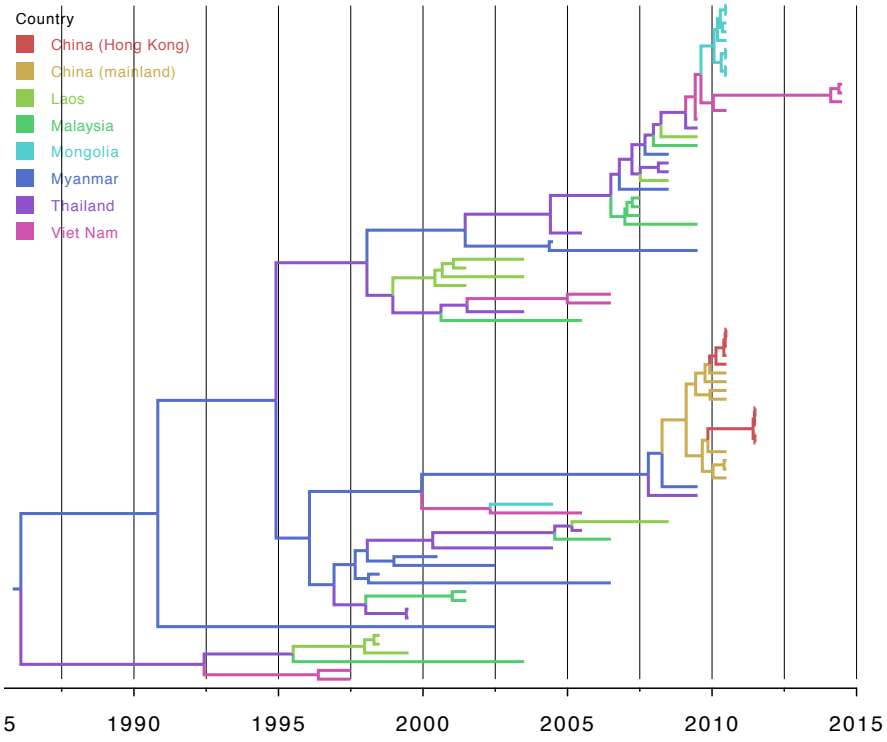
(l) GLM results, SEA6



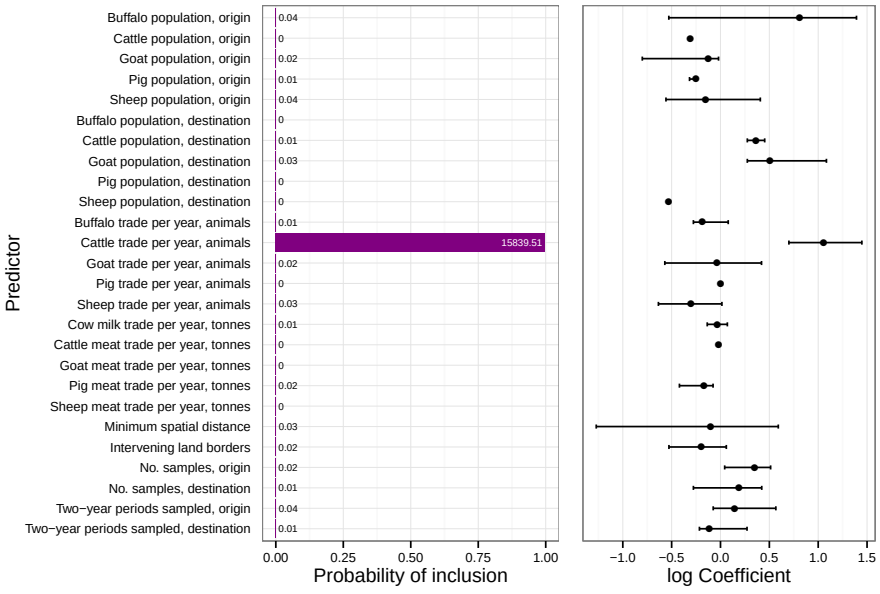
(m) MCC tree, SEA7



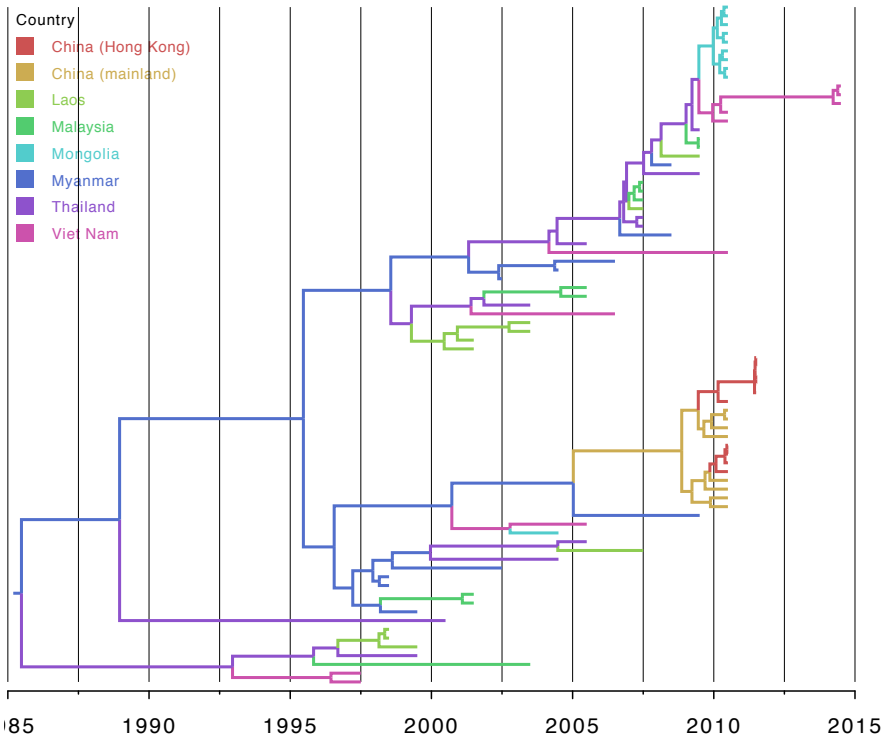
(n) GLM results, SEA7



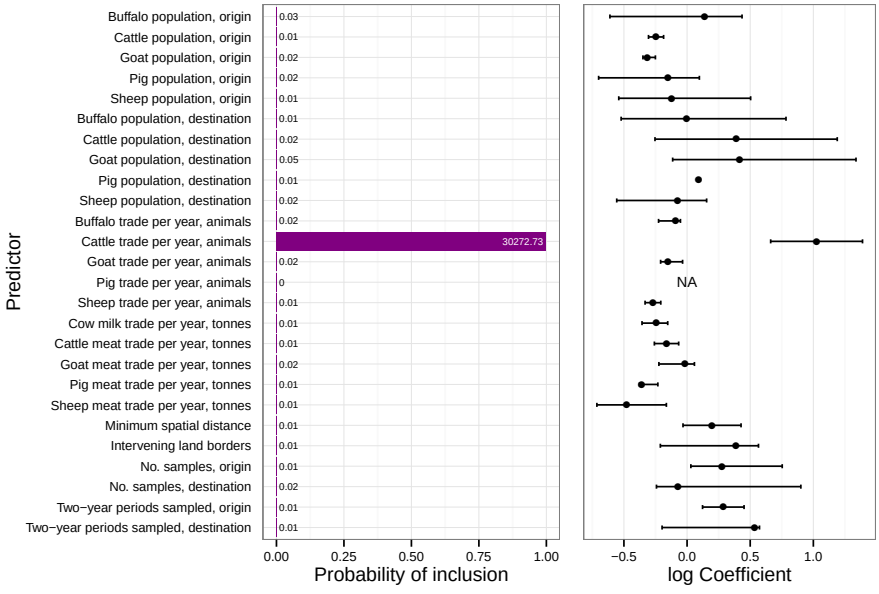
(o) MCC tree, SEA8



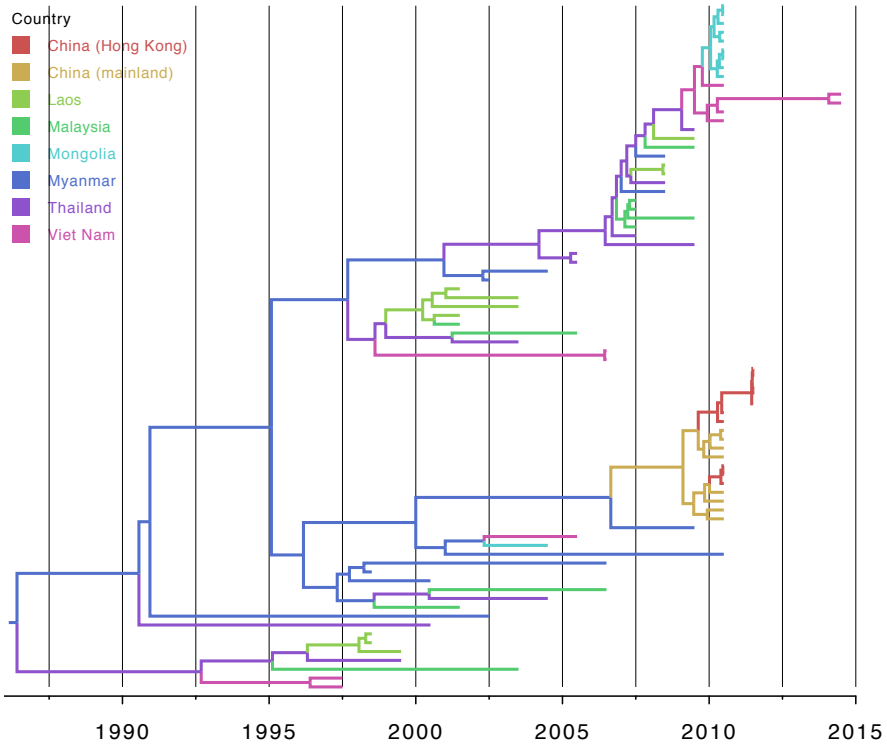
(p) GLM results, SEA8



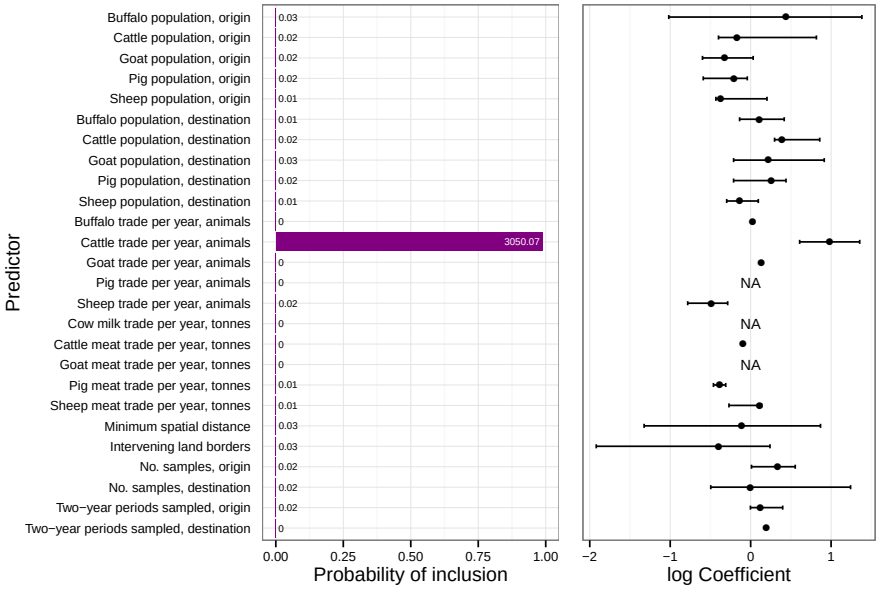
(q) MCC tree, SEA9



(r) GLM results, SEA9



(s) MCC tree, SEA10



(t) GLM results, SEA10

Figure B.7: Maximum clade credibility trees and GLM predictor results for each sampling replicate of the analysis of the full serotype. Branches in the tree are coloured by most probable location.

Replicate	Host	Mean $N_e\tau$	Median $N_e\tau$	95% HPD interval
SEAH1	Cattle	2852.03	40.67	0.76-4305.64
	Pigs	1.12	0.94	0.28-2.39
	B. bubalis	7.57	5.75	0.75-11.09
SEAH2	Cattle	8147.59	81.12	1.03-6939.96
	Pigs	0.87	0.77	0.30-1.67
	B. bubalis	5.38	3.21	0.37-12.24
SEAH3	Cattle	5273.32	47.75	1.53-3625.25
	Pigs	1.60	1.35	0.36-3.46
	B. bubalis	17.70	3.55	0.36-12.55
SEAH4	Cattle	5375.38	23.90	0.65-325.14
	Pigs	2.19	1.85	0.53-4.62
	B. bubalis	16.11	3.06	0.31-15.49
SEAH5	Cattle	3412.90	43.42	1.21-3674.83
	Pigs	1.17	1.06	0.40-2.16
	B. bubalis	9.53	3.86	0.39-13.45
SEAH6	Cattle	18968.56	45.85	0.60-3866.45
	Pigs	1.42	1.28	0.47-2.72
	B. bubalis	12.06	3.63	0.47-11.81
SEAH7	Cattle	5349.69	91.84	0.79-7005.47
	Pigs	5.10	4.05	0.84-11.78
	B. bubalis	6.93	1.95	0.25-7.91
SEAH8	Cattle	9687.74	164.85	1.74-11957.76
	Pigs	2.21	1.93	0.66-4.46
	B. bubalis	3.15	2.14	0.43-8.68
SEAH9	Cattle	3579.39	28.98	0.72-2468.09
	Pigs	1.16	0.97	0.26-2.41
	B. bubalis	45.36	5.51	0.39-15.73
SEAH10	Cattle	2786.11	24.93	0.42-2854.42
	Pigs	0.85	0.74	0.20-1.71
	B. bubalis	5.56	4.91	0.47-10.81

Table B.1: Summary of posterior distribution for effective population sizes of host demes, BASTA analysis of topotype SEA.

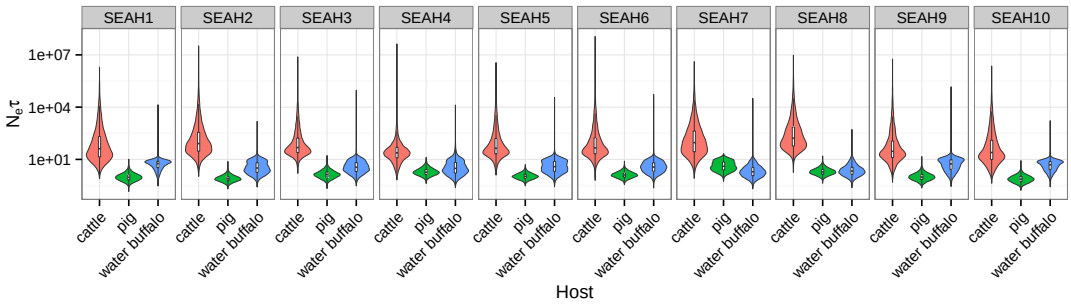


Figure B.8: Estimated posterior distributions for host deme sizes in the BASTA analysis, across ten replicates of the sampling scheme.

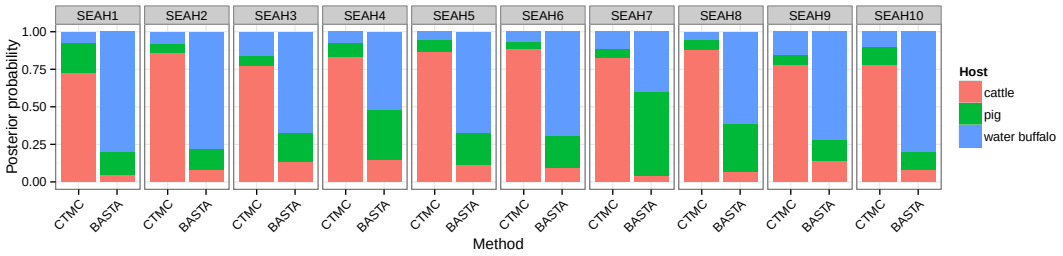
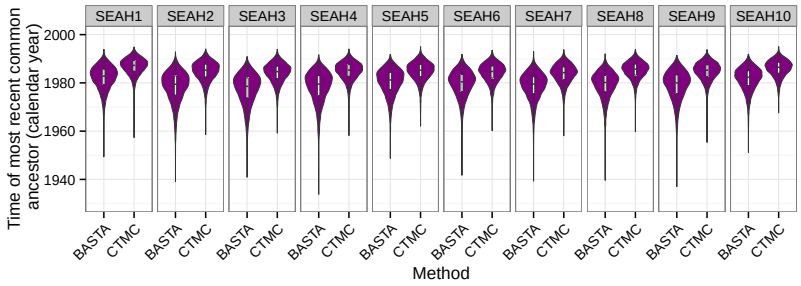


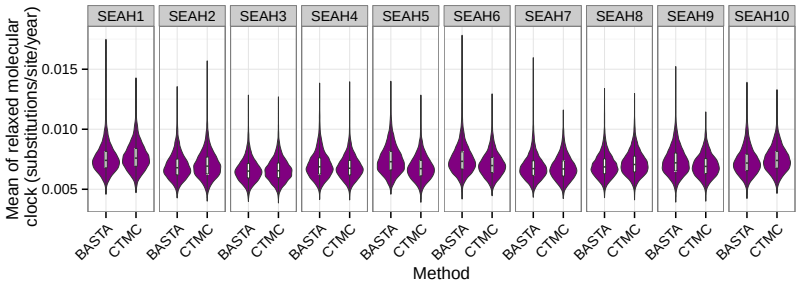
Figure B.9: Estimated posterior distributions for the host species of the lineage at the root of the phylogeny, comparing CTMC discrete traits and BASTA analysis, across ten replicates of the sampling scheme.

rate, although BASTA does estimate somewhat larger standard deviations for this rate.

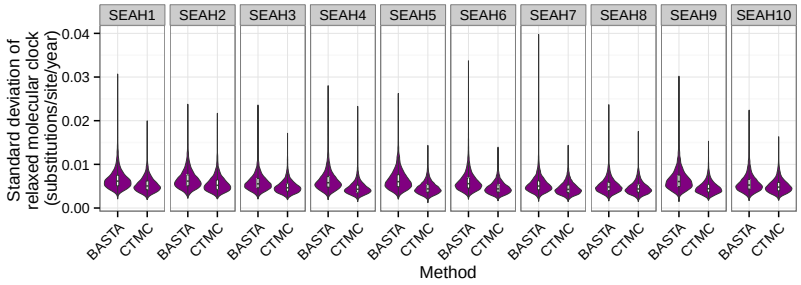
In figure B.11, the posterior median estimates for each rate of transition are compared, by replicate. Since CTMC rates are forwards in time and BASTA backwards, the absolute sizes of these numbers are not directly comparable. Estimates were always considerably higher for BASTA, generally by a factor of around ten. All replicates except one (SEAH1) showed clear correlation between the estimates from each method.



(a) TMRCA



(b) Molecular clock mean



(c) Molecular clock standard deviation

Figure B.10: Estimated posterior distributions for (a) the TMRCA of topotype SEA and the mean (b) and standard deviation (c) of the lognormal distribution governing molecular clock rates in this topotype. Each violin represents a different sampling replicate.

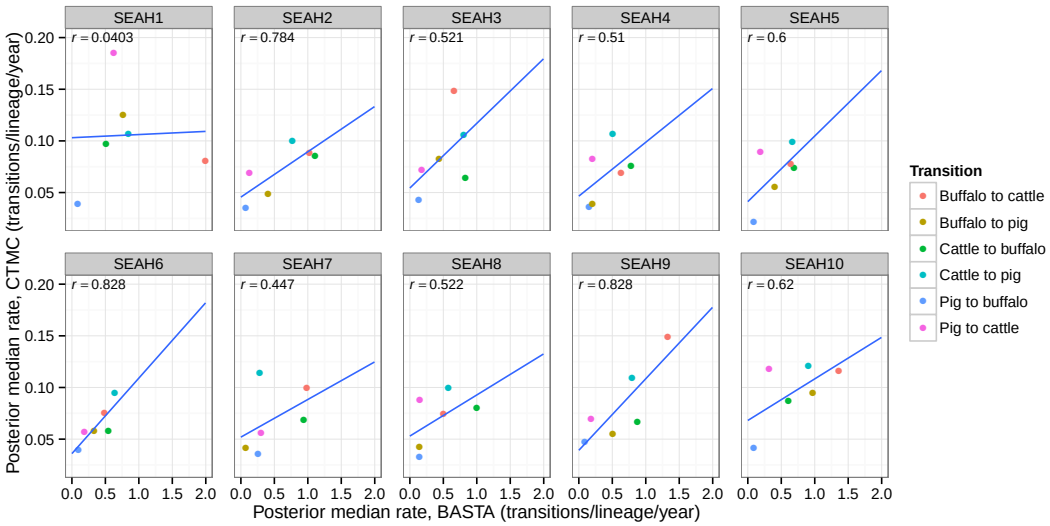


Figure B.11: The posterior median estimates for the rate of each host-to-host transition, from ten sampling replicates, topotype SEA. Blue lines were fit by simple linear regression and plots are labelled with the correlation coefficient.

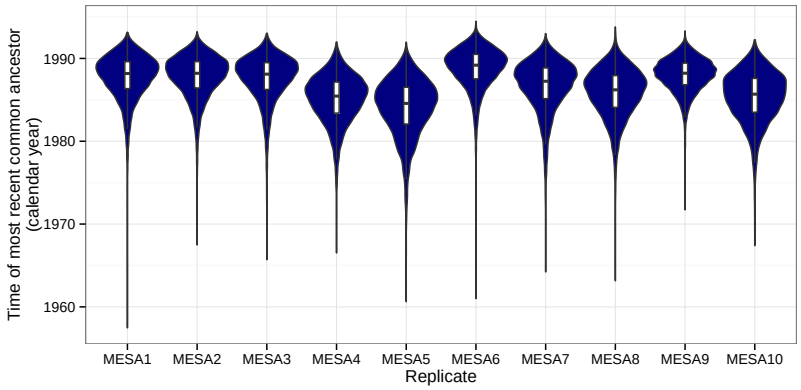
B.2.3 Topotype ME-SA

The ten replicates of the phylogeography sampling scheme are designated MESA1 to MESA10 and the ten of the host species sampling scheme MESAH1 to MESAH10. MESA1 and MESAH1 are presented in chapter 4. Figure B.12 depicts the posterior distributions for the TMRCA of all the included sequences and the parameters of the molecular clock. Figure B.13 displays the reconstructed skygrid plots, figure B.14 the posterior distributions for the geographical location of the root of the tree, and figure B.15 the MCC trees and GLM predictor results. There is rather more variation in these estimates than there was for SEA. The earliest posterior median TMRCA was July 1984 (MESA5, January 1977-February 1990) and the latest March 1989 (MESAS6, September 1983-January 1993). Mean rates for the molecular clock ranged from 6.23×10^{-3} (MESA9, $5.11 \times 10^{-3} - 7.57 \times 10^{-3}$) to 7.16×10^{-3} (MESA10, $5.86 \times 10^{-3} - 8.71 \times 10^{-3}$) substitutions per site per year. Turkey is by some distance the most probably root location in eight out of

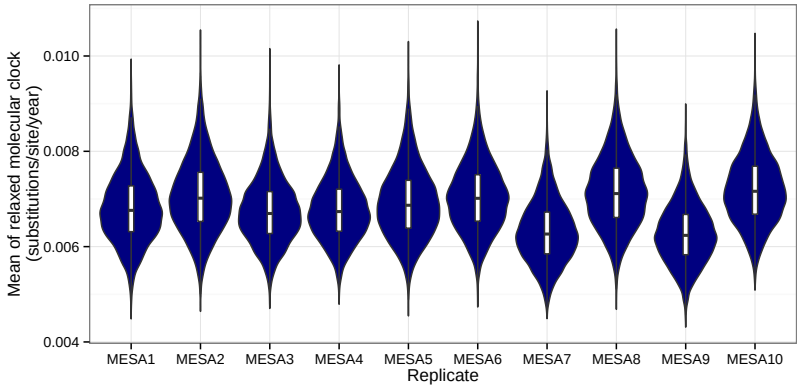
ten replicates, but the other two show considerable uncertainty, with substantial support for Pakistan, Iran, India, Nepal and Bhutan. While the inclusion of minimum geographical distance as a predictor of movement is always strongly supported, a variety of different predictors regarding the nature of the included samples were in different replicates with varying Bayes Factors (BFs). The only epidemiological predictor that ever obtains a BF greater than 1 is trade in cow milk in MESA2.

The estimated posterior distributions of the effective population sizes of the demes comprising viruses infecting cattle, pigs and buffalo from the BASTA analysis are summarised in figure B.16. Once again numerical results are also given in table B.2. These results lack consistency. The cattle deme generally had the largest median size in seven out of ten replicates, and the buffalo deme in the remaining two. The enormous variance in the size of the cattle deme seen in SEA was also seen here in all but one replicate; but in this case it was also frequently seen for the other demes too.

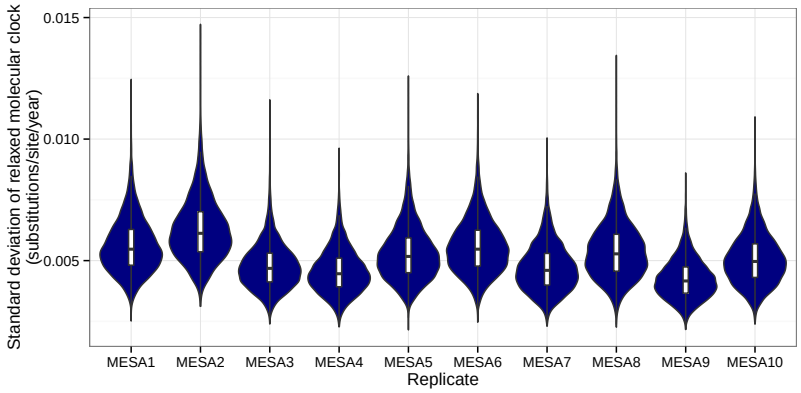
Figure B.17 shows the posterior distributions of the host assigned to the root of the phylogeny. These again differ greatly (except for MESAH6), with CTMC always preferring cattle and BASTA generally buffalo but sometimes cattle. Posterior median TMRCAs of all sequences were again rather earlier for BASTA (figure B.18). CTMC figures, which were more consistent than those from the phylogeography analysis and universally later, ranged from October 1990 (MESAH6, December 1986-July 1993) to August 1992 (MESAH4, May 1989-December 1994). BASTA figures were from February 1986 (MESAH2, February 1973-May 1992) to February 1989 (MESAH5, December 1983-March 1993). Despite estimating earlier TMRCAs, BASTA tended to give faster mean molecular clock rates. CTMC mean rates ranged from 6.39×10^{-3} (MESAH3, $4.51 \times 10^{-3} - 8.8 \times 10^{-3}$) to 0.0102 (MESAH4, $7.32 \times 10^{-3} - 0.014$) and BASTA from 7.48×10^{-3} (MESAH3, $5.43 \times 10^{-3} - 0.0104$) to 0.0133 (MESAH5, $6.92 \times 10^{-3} - 0.0172$) substitutions per site per year. Finally,



(a) TMRCA



(b) Molecular clock mean



(c) Molecular clock standard deviation

Figure B.12: Violin plots for the posterior distribution of a) the TMRCA of all sequences, b) the mean and c) the standard deviation of the uncorrelated lognormal molecular clock in each sampling replicate of the analysis of the ME-SA topotype.

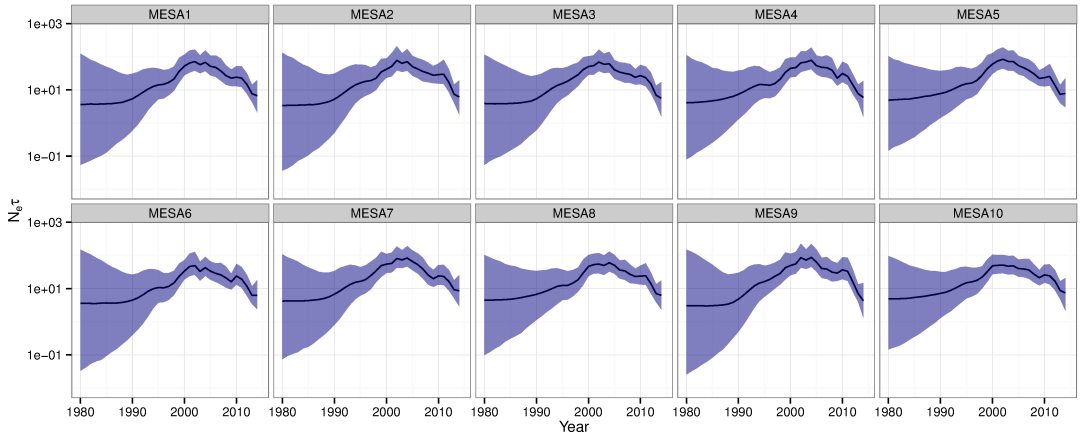


Figure B.13: Reconstructed skygrid plots for each sampling replicate of the analysis of the ME-SA topotype. The black line is the median effective population size and the grey area the 95% highest posterior density region.

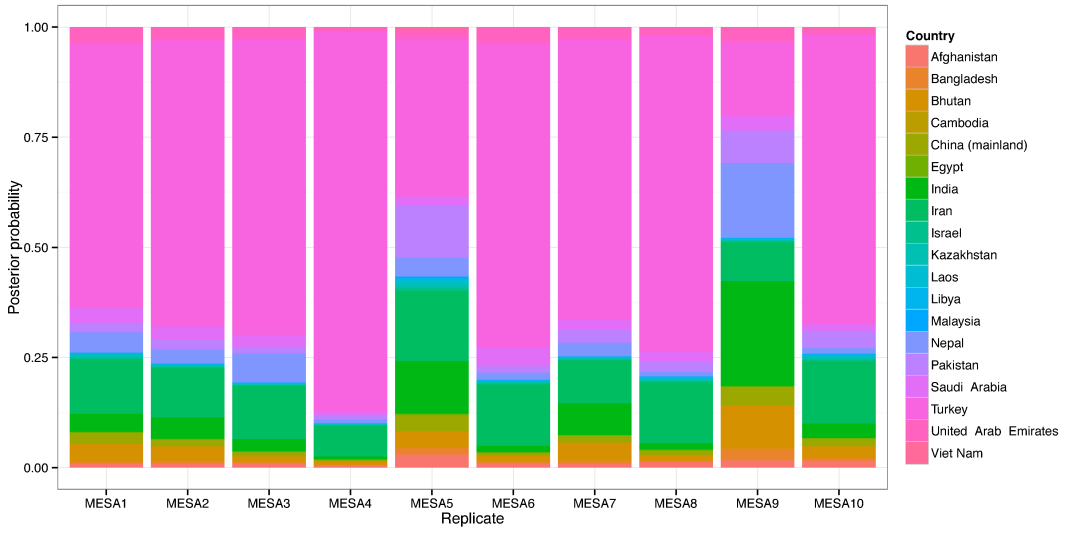
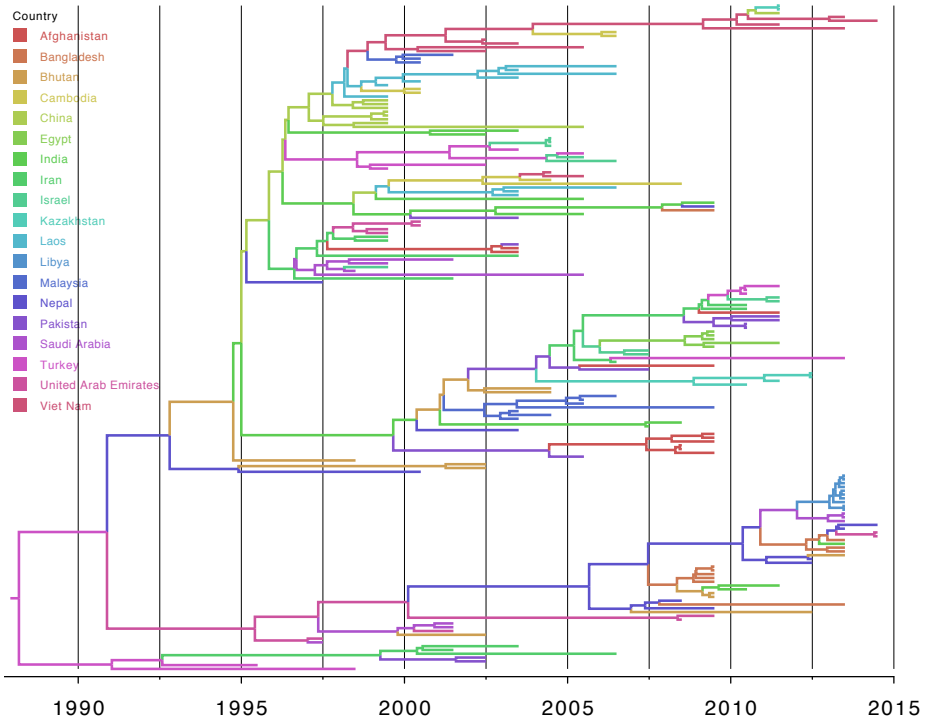
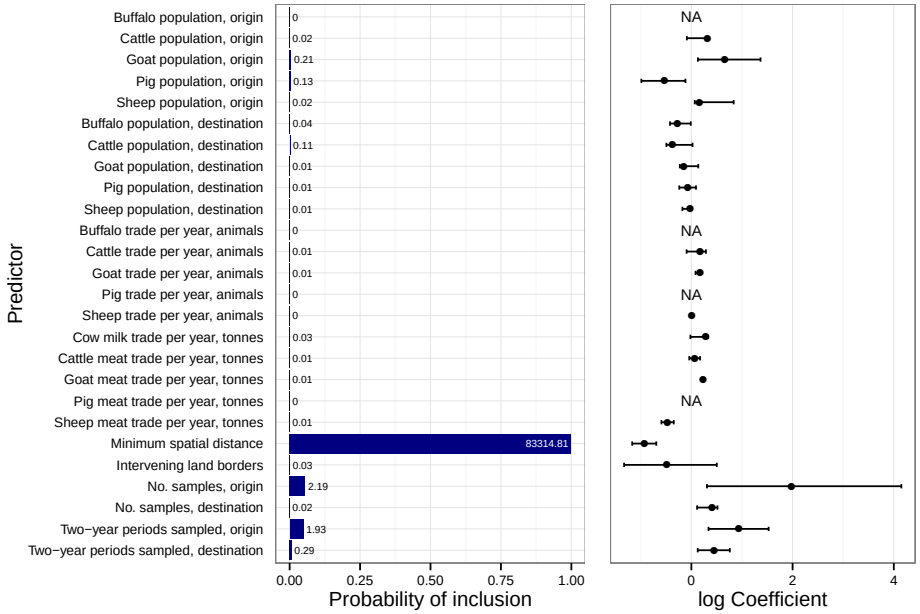


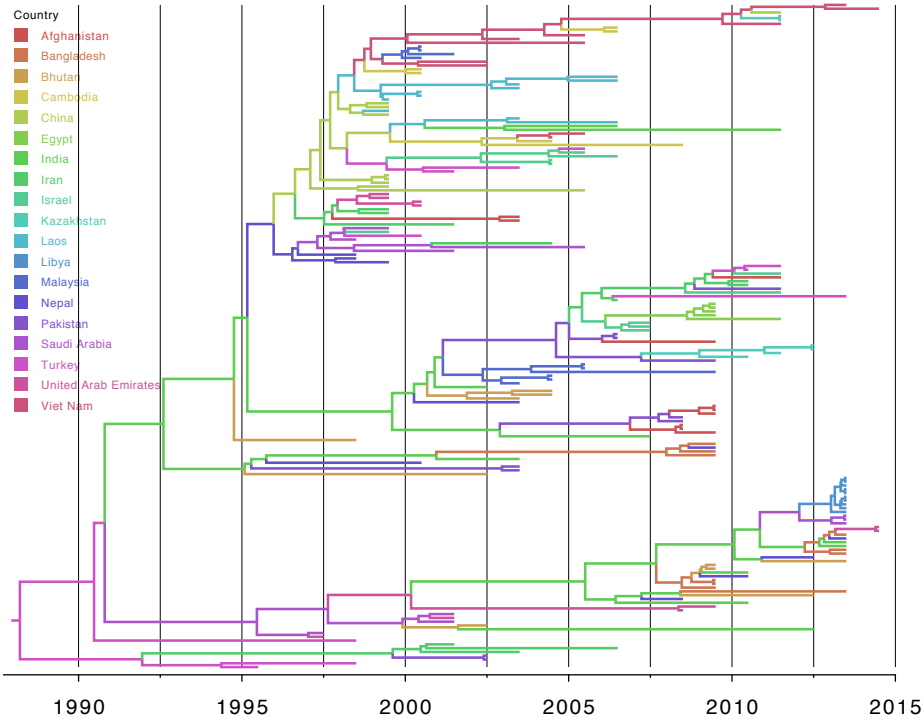
Figure B.14: Posterior distributions for the geographical location of the root node for each sampling replicate of the analysis of the ME-SA topotype.



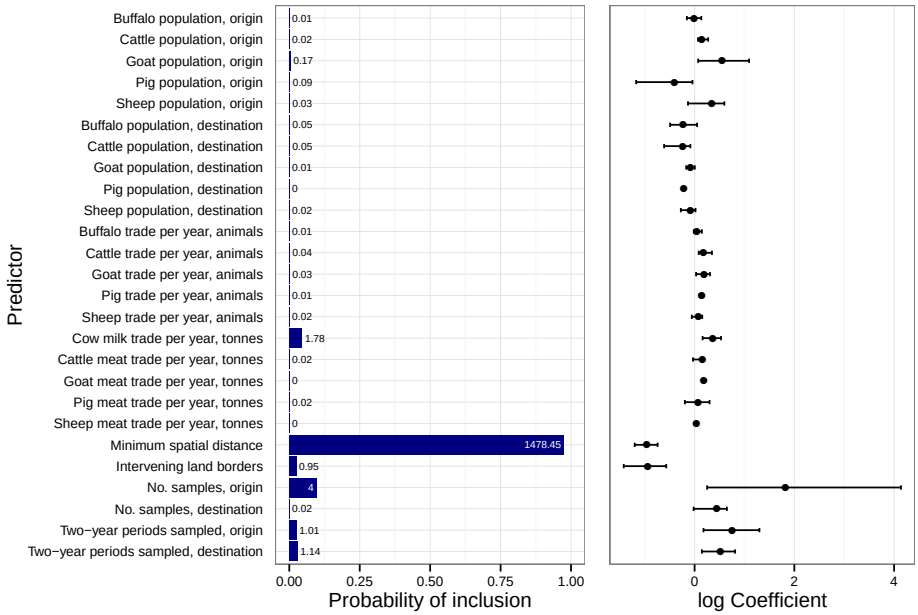
(a) MCC tree, MESA1



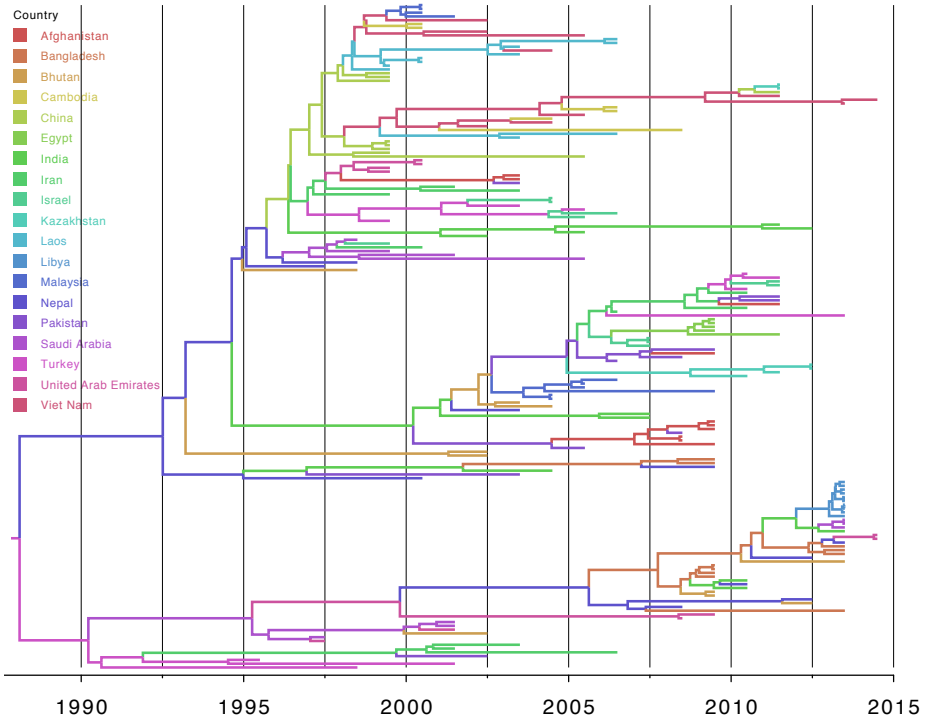
(b) GLM results, MESA1



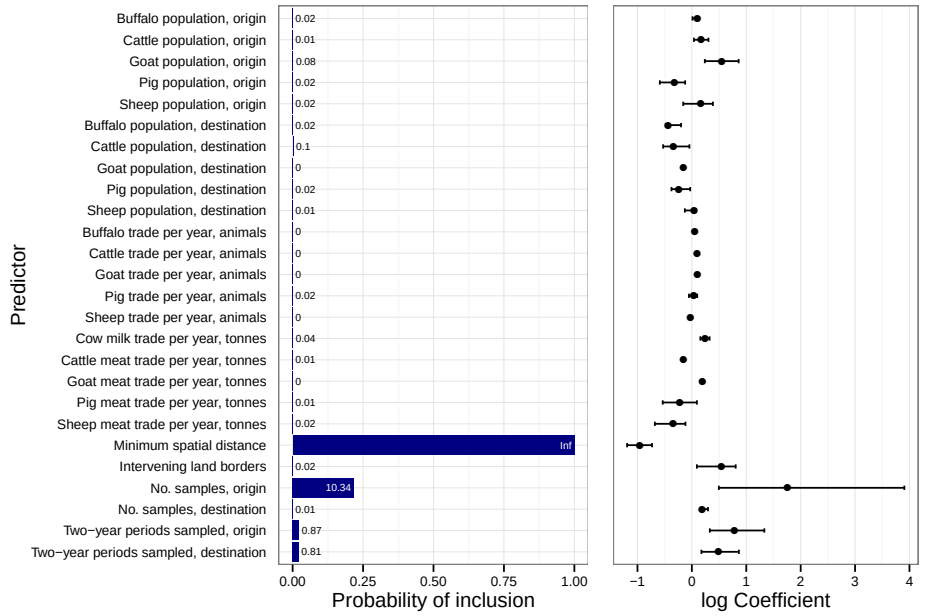
(c) MCC tree, MESA2



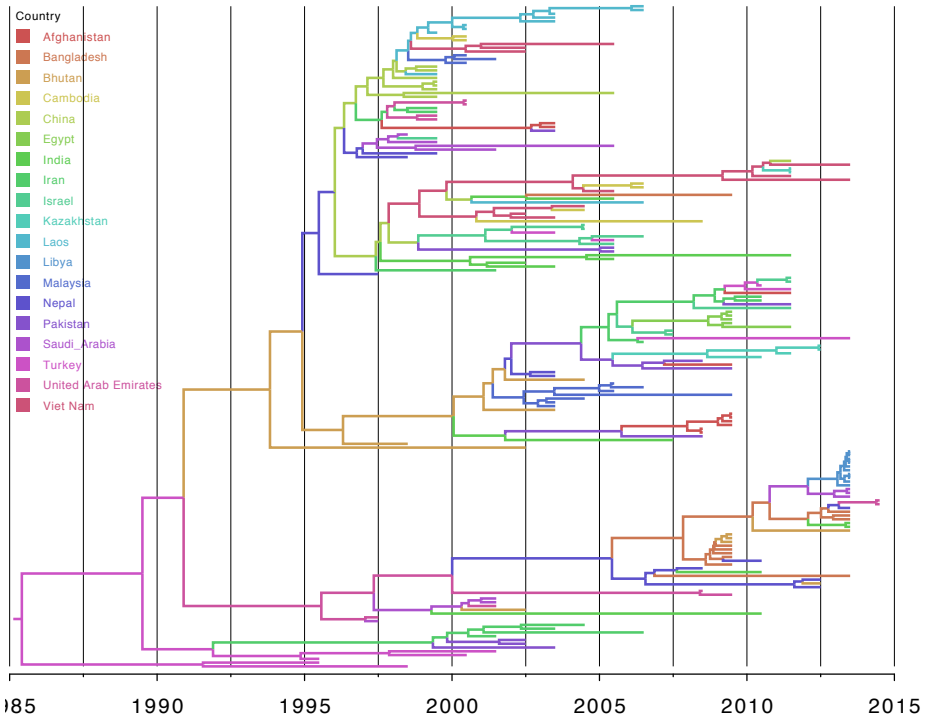
(d) GLM results, MESA2



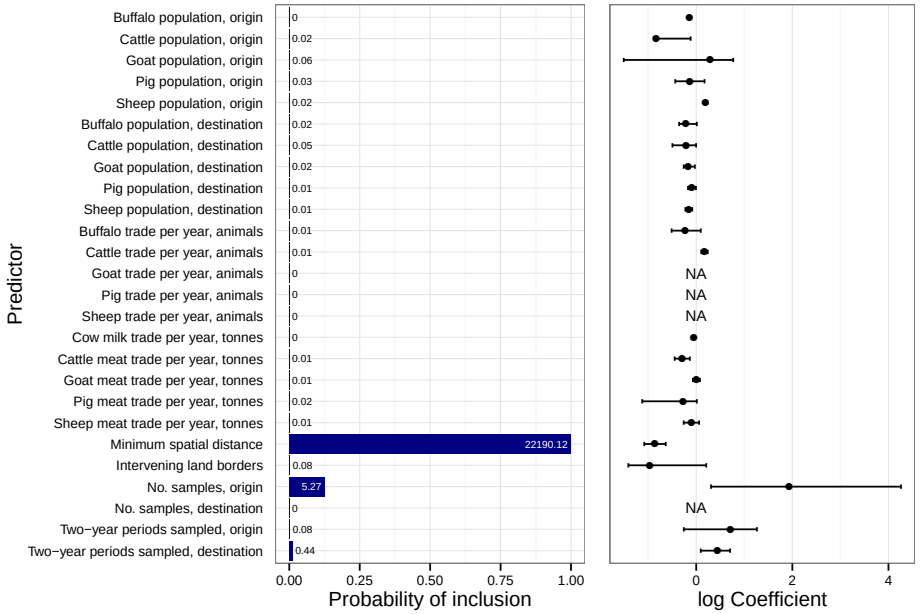
(e) MCC tree, MESA3



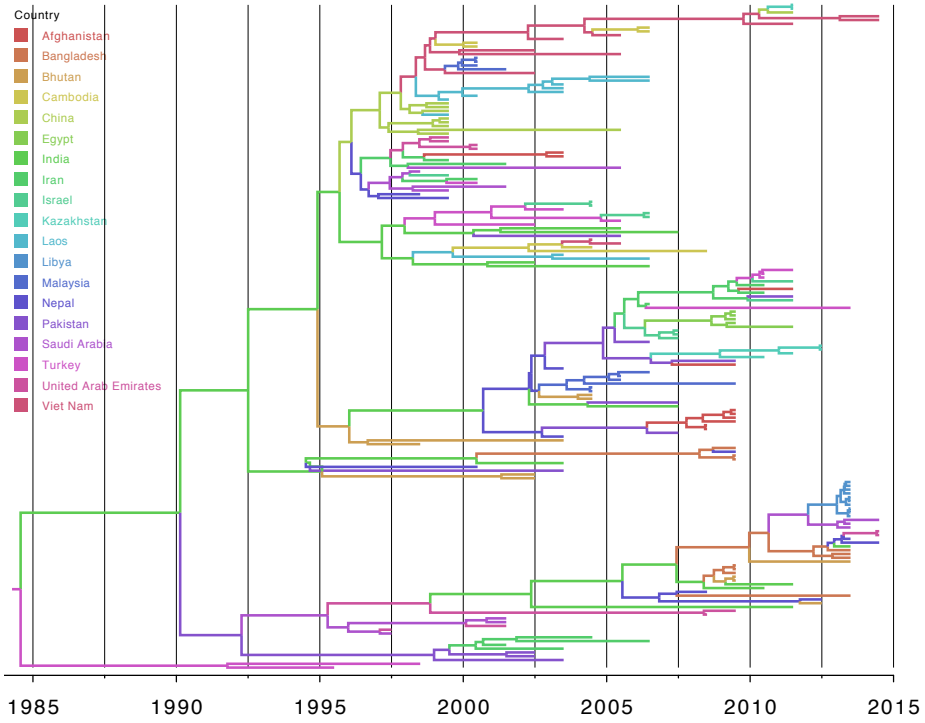
(f) GLM results, MESA3



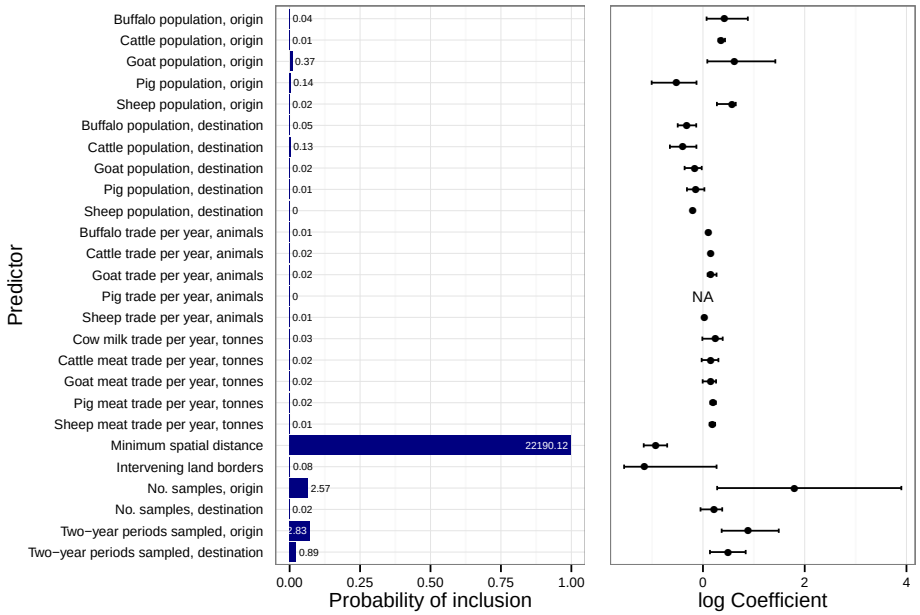
(g) MCC tree, MESA4



(h) GLM results, MESA4

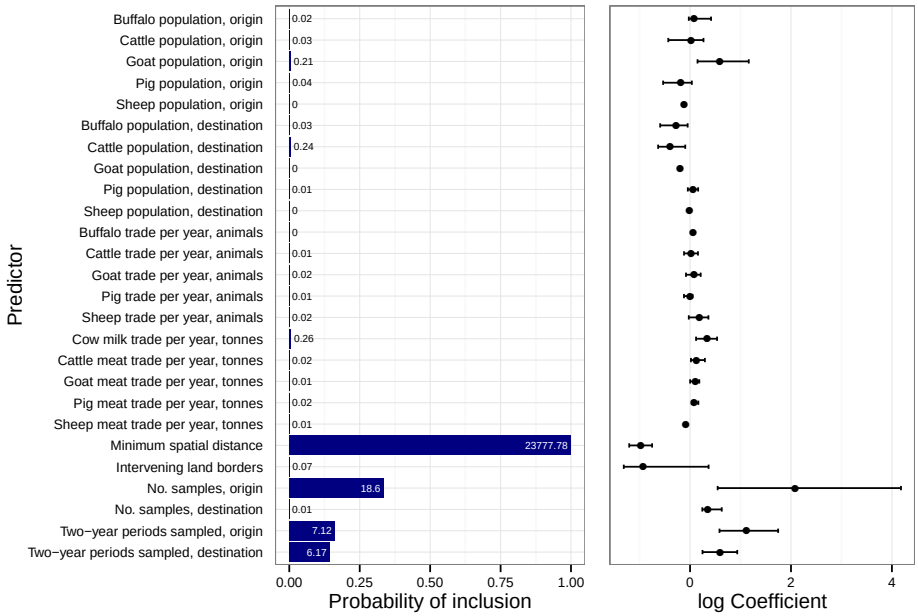


(i) MCC tree, MESA5

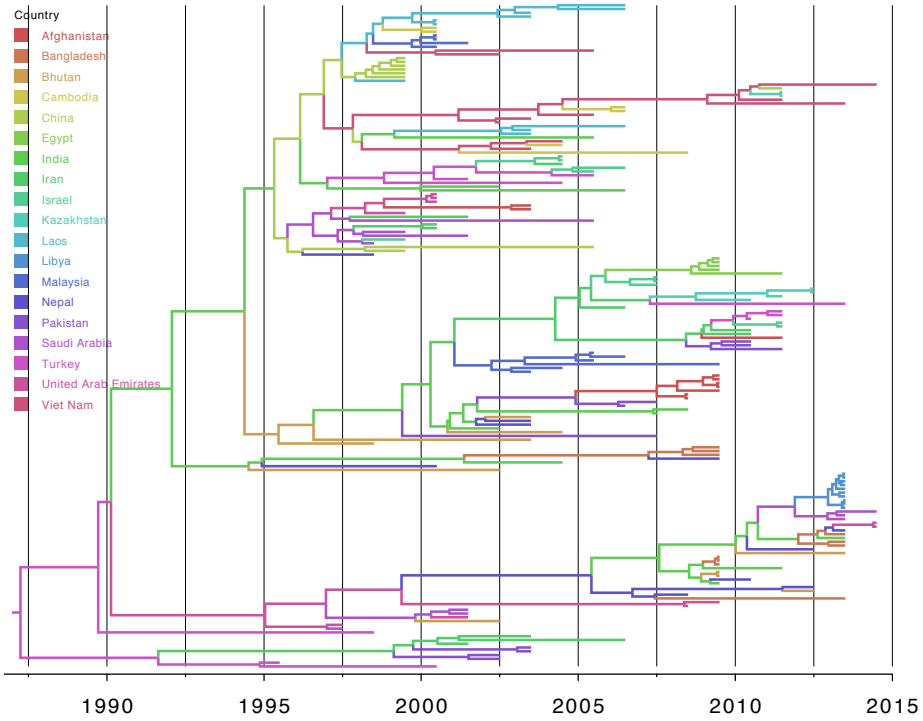


(j) GLM results, MESA5

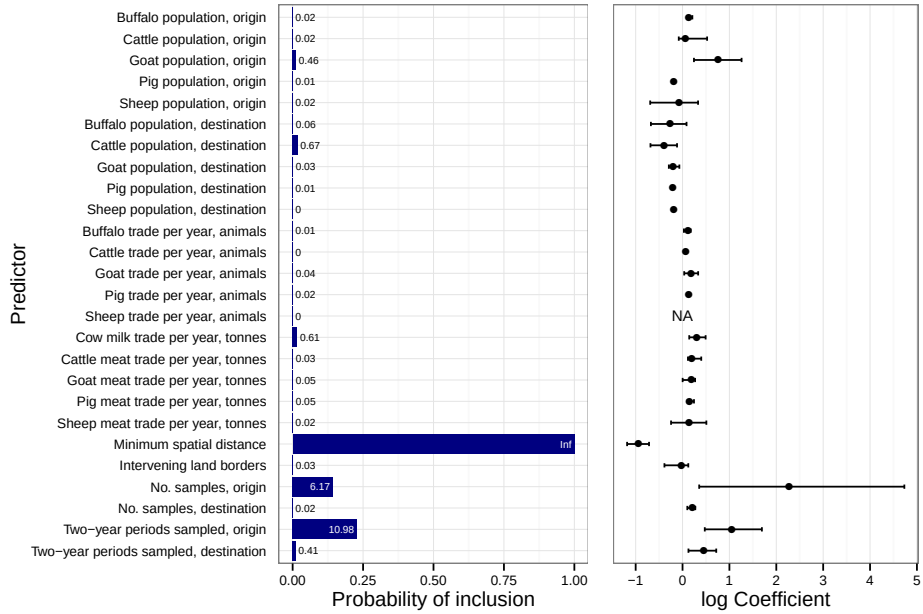
(k) MCC tree, MESA6



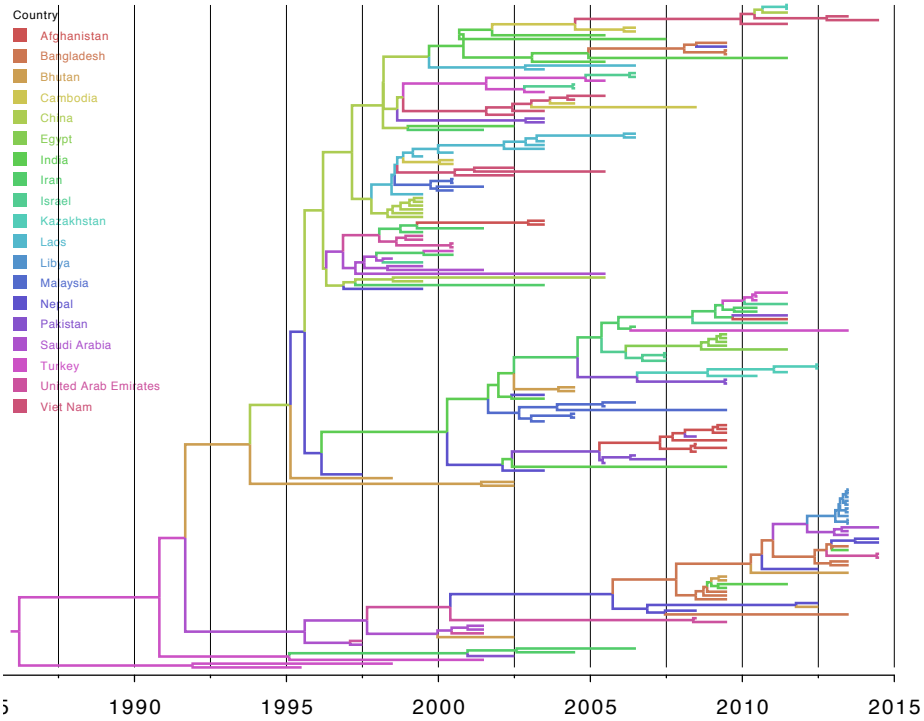
(l) GLM results, MESA6



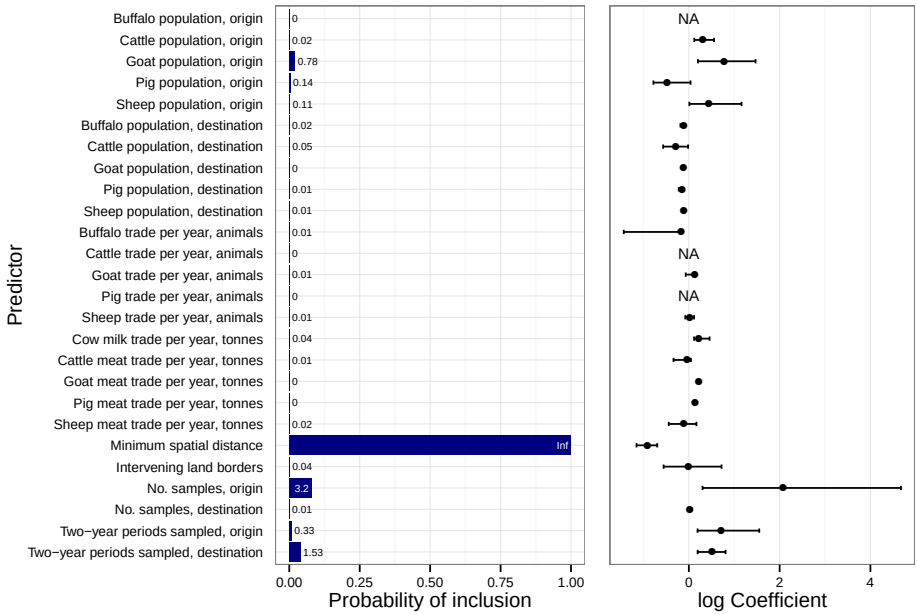
(m) MCC tree, MESA7



(n) GLM results, MESA7

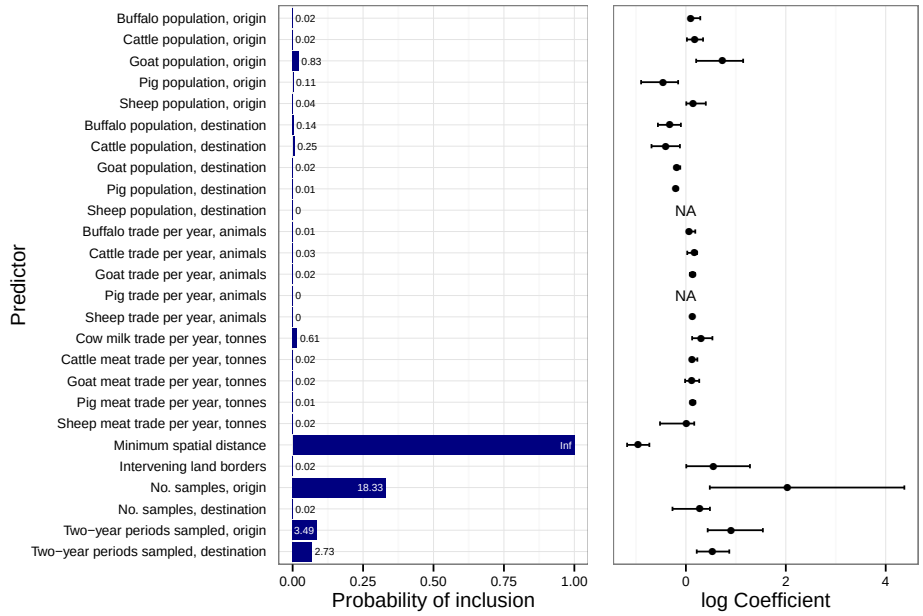


(o) MCC tree, MESA8

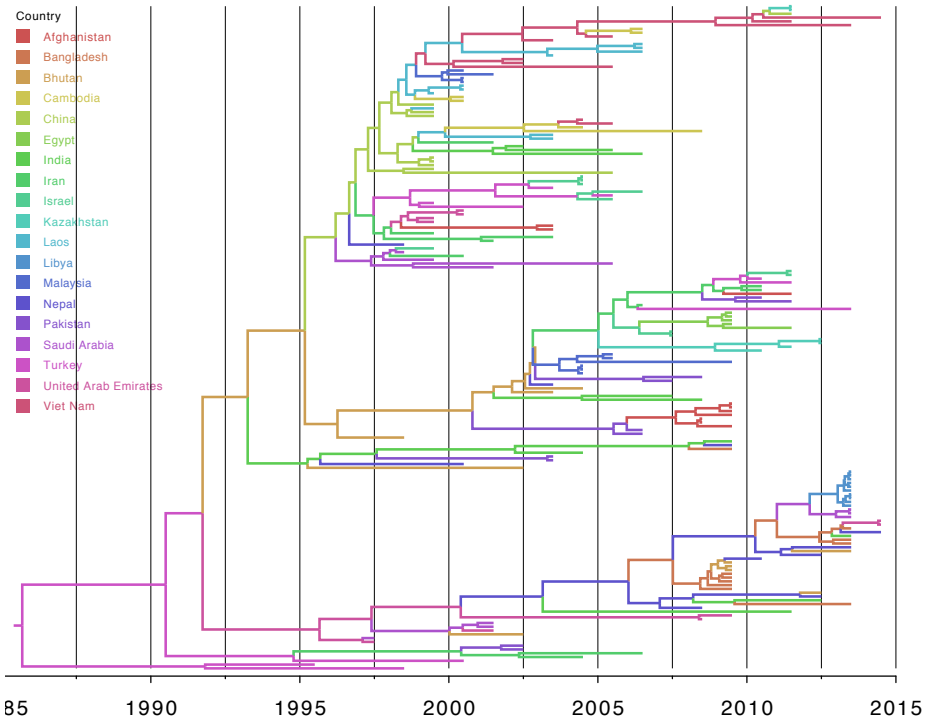


(p) GLM results, MESA8

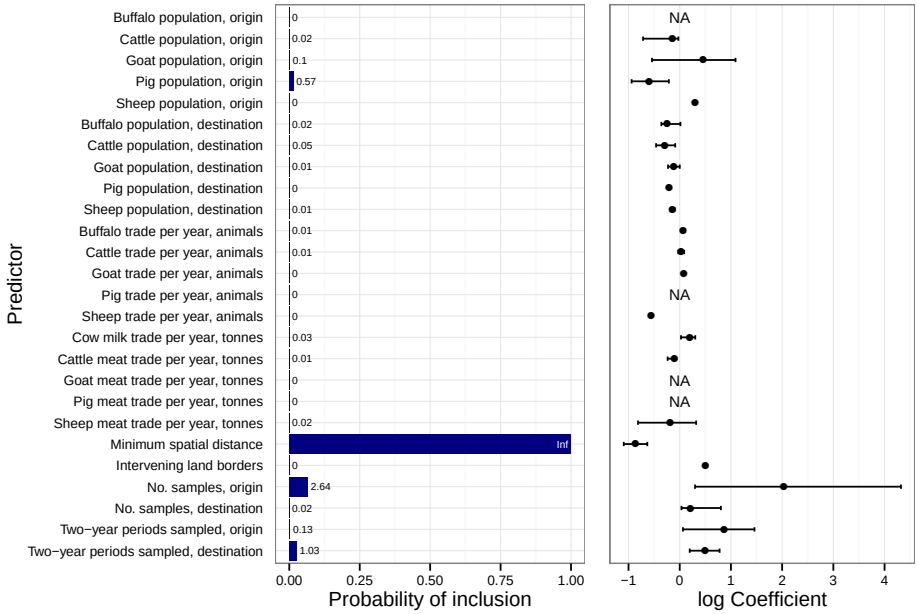
(q) MCC tree, MESA9



(r) GLM results, MESA9



(s) MCC tree, MESA10



(t) GLM results, MESA10

Figure B.15: Maximum clade credibility trees and GLM predictor results for each sampling replicate of the analysis of topotype ME-SA. Branches in the tree are coloured by most probable location

Replicate	Host	Mean $N_e\tau$	Median $N_e\tau$	95% HPD interval
MESAH1	Cattle	10050.68	20.68	0.17-1455.84
	Pigs	486.60	2.65	0.20-73.57
	Sheep	415.07	0.57	0.17-1.35
	B. bubalis	13.71	13.66	0.48-26.61
MESAH2	Cattle	2979.02	25.54	0.23-3558.40
	Pigs	131.56	1.64	0.26-9.61
	Sheep	8.03	0.56	0.21-1.12
	B. bubalis	19.89	19.31	12.15-28.22
MESAH3	Cattle	4907.00	46.40	0.45-992.86
	Pigs	186.31	1.34	0.20-34.00
	Sheep	590.54	0.69	0.14-179.86
	B. bubalis	1362.69	7.17	0.69-34.96
MESAH4	Cattle	951.01	0.71	0.12-39.27
	Pigs	102.84	0.95	0.19-3.43
	Sheep	5258.38	1.45	0.13-1172.41
	B. bubalis	17.82	17.54	10.83-25.91
MESAH5	Cattle	2080.59	8.67	0.17-1532.24
	Pigs	117.31	0.79	0.15-13.88
	Sheep	928.75	0.56	0.13-133.08
	B. bubalis	14.72	13.32	8.03-19.53
MESAH6	Cattle	26.27	25.78	16.94-37.11
	Pigs	3269.53	1.19	0.17-889.26
	Sheep	1778.75	0.75	0.12-800.48
	B. bubalis	9.81	1.36	0.39-3.70
MESAH7	Cattle	2395.20	50.29	1.09-2863.36
	Pigs	341.83	0.73	0.11-46.19
	Sheep	5575.90	0.59	0.12-323.33
	B. bubalis	401.86	2.40	0.32-18.29
MESAH8	Cattle	9541.57	20.56	0.96-426.60
	Pigs	51.61	1.51	0.21-6.06
	Sheep	301.82	0.59	0.17-1.55
	B. bubalis	7.00	2.41	0.38-14.98
MESAH9	Cattle	1587.50	6.99	0.14-1483.59
	Pigs	6.50	1.67	0.26-10.05
	Sheep	68.12	0.56	0.11-3.88
	B. bubalis	20.44	17.13	2.57-25.99
MESAH10	Cattle	447.08	28.01	0.89-140.10
	Pigs	130.59	1.01	0.17-20.98
	Sheep	268.13	0.58	0.13-138.73
	B. bubalis	3.77	1.58	0.26-18.90

Table B.2: Summary of posterior distribution for effective population sizes of host demes, BASTA analysis of topotype ME-SA

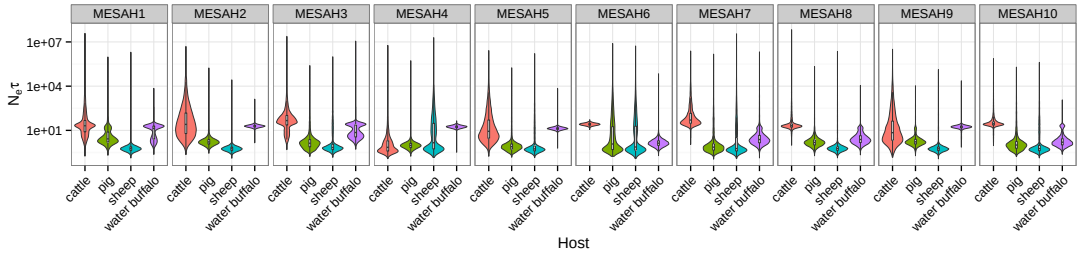


Figure B.16: Estimated posterior distributions for host deme sizes in the BASTA analysis, across ten replicates of the sampling scheme, topotype ME-SA.

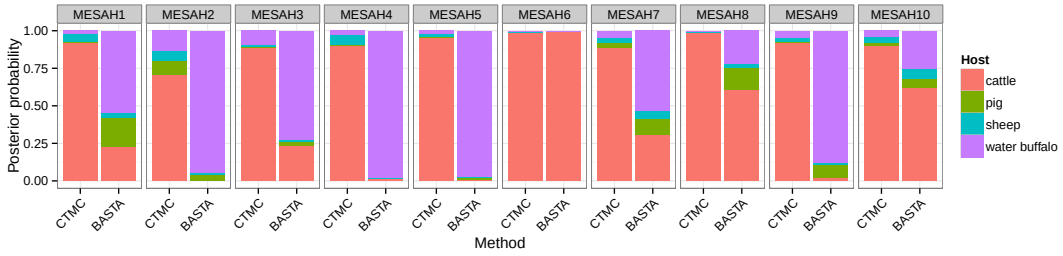
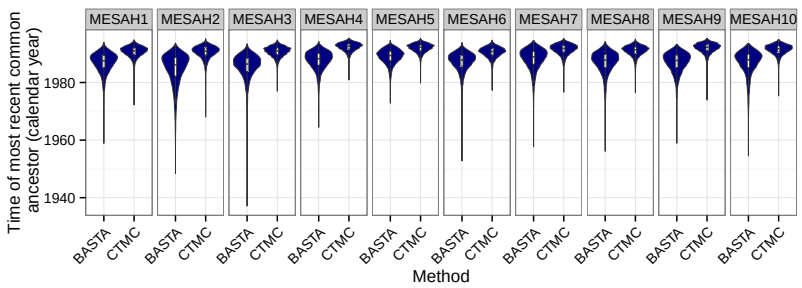
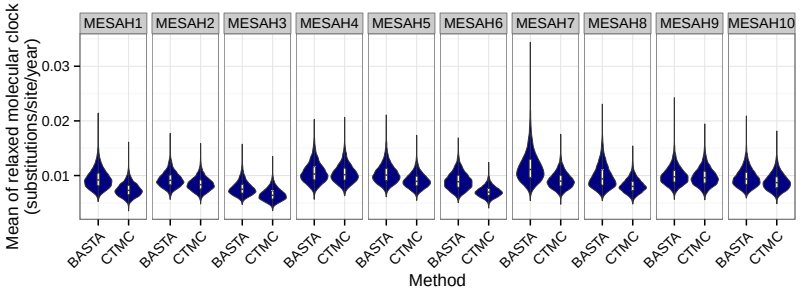


Figure B.17: Estimated posterior distributions for the host species of the lineage at the root of the phylogeny, comparing CTMC discrete traits and BASTA analysis, across ten replicates of the sampling scheme.

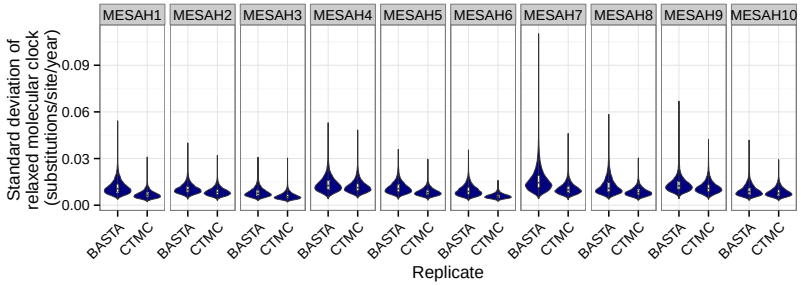
while transition rate estimates from BASTA were again around ten times higher than they were from CTMC (figure B.19), estimates from the two methods were only well-correlated with each other for one replicate (MESAH8) and for several replicates (MESAH1, MESAH4, MESAH9) there was little suggestion that any relationship existed.



(a) TMRCA



(b) Molecular clock mean



(c) Molecular clock standard deviation

Figure B.18: Estimated posterior distributions for (a) the TMRCA of topotype ME-SA and the mean (b) and standard deviation (c) of the lognormal distribution governing molecular clock rates in this topotype. Each violin represents a different sampling replicate.

B.3 Discussion

The principal point of interest in the reconstruction of the full serotype concerns replicate ALL10, which is unique in several ways: it estimated a much earlier TMRCA, it detected a rate change point for the SEA and African clades, and it was the one replicate for which the posterior probability that no contemporary Euro-SA(1) sequences were descended from O_1 Campos/58 vaccines was higher than the probability that any were. It could be argued that this replicate presents a more plausible scenario than the others. The early TMRCA is consistent with the 1870 date for introduction to South America [125], and there is a difference between tip rates and deep tree rates for almost every topotype, which would be expected (see section 4.4). Nevertheless, only one out of ten replicates showed this pattern and the posterior median rate of 5×10^{-3} substitutions per site per year for SEA differs considerably from that found from the analysis of that topotype alone (lognormally distributed with a median of 6.75×10^{-3}).

That, in general, there was more variation between ME-SA replicates than SEA replicates is likely to be simply because the overall pool of sequences to draw from was considerably larger in the latter case. It is comforting that there were no vast differences between sampling replicates for most outputs: HPD intervals for numerical parameters always overlapped with each other, and while the set of supported predictors was subject to some variation, that would not change the epidemiological interpretation if the threshold for a well-supported predictor was a BF of 3. The sample-related predictors are best regarded as nuisance parameters, and their inclusion in the model makes it possible to control for them.

The CTMC approach, in which host species is treated as a characteristic of a particular virus that evolves independently to any population structure, is a major simplification, but what I have shown here when comparing it to BASTA is that this may still be preferable to imposing an overly simplistic population structure.

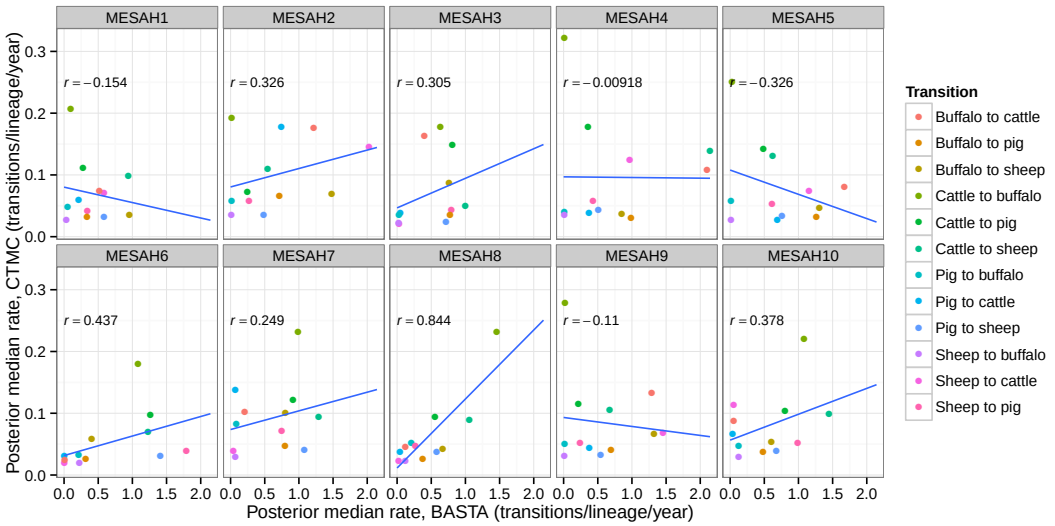


Figure B.19: Comparison of posterior median estimates for the rate of each host-to-host transition from CTMC and BASTA analyses, from ten sampling replicates, topotype ME-SA. Blue lines were fit by simple linear regression and plots are labelled with the correlation coefficient.

Given the known epidemiology of the virus, the CTMC suggestion that the root host for both topotypes was most likely to be cattle is much more convincing than the BASTA suggestion that it was *B. bubalis*. As BASTA cannot currently reconstruct ancestral states on a tree, it is not entirely clear why the latter inference was made, but it is probably due to deme sizes. That the cattle deme is usually inferred to be much larger than the others, albeit with estimates of its size having an enormous variance, is realistic. But this size means that the coalescence of two particular lineages in this deme (which is, in the model, a freely-mixing population of every cattle lineage on the planet) is much less likely than the coalescence of two in another, smaller deme. If the population structure was further subdivided (probably geographically), this effect might disappear. Future analyses using this method should take this into account. The lack of correlation between transition rates inferred by each method for ME-SA is quite concerning, suggesting that the two methods may be inferring rather different transmission structures, but full

investigation of the reasons for this will have to await more sophisticated analysis tools for BASTA output.

The BASTA results also show much more between-replicate variation than the CTMC results do, in terms of both deme sizes and the reconstructed root host. Since one of the key arguments for the use of a structured coalescent was to eliminate sampling effects introduced by the CTMC approach, possible reasons for this may warrant further investigation. A difference between estimated root heights would be expected as a result of structuring the model, as a delay in the time of the final coalescence (in backwards time) is to be expected if two lineages have to wait until have migrated into the same deme before they can coalesce. Nevertheless, I would hesitate in favouring the BASTA numbers for the reasons outlined above; the root host reconstruction appears to be influenced by an unrealistic population model, and as the deeper roots are a consequence of that model, this may also be less than trustworthy.

Appendix C

Related publication

Chapter 2 was published in *mBio* in 2013 [59]. The journal version is reproduced here.

RESEARCH ARTICLE

Reconstructing Geographical Movements and Host Species Transitions of Foot-and-Mouth Disease Virus Serotype SAT 2

Matthew D. Hall,[a,b] Nick J. Knowles,[c] Jemma Wadsworth,[c] Andrew Rambaut,[a,b,d] Mark E. J. Woolhouse[a,b]

Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom[a]; Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, United Kingdom[b]; The Pirbright Institute, Pirbright, Surrey, United Kingdom[c]; Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA[d]

ABSTRACT Of the three foot-and-mouth-disease virus SAT serotypes mainly confined to sub-Saharan Africa, SAT 2 is the strain most often recorded in domestic animals and has caused outbreaks in North Africa and the Middle East six times in the last 25 years, with three apparently separate events occurring in 2012. This study updates the picture of SAT 2 phylogenetics by using all available sequences for the VP1 section of the genome available at the time of writing and uses phylogeographic methods to trace the origin of all outbreaks occurring north of the Sahara since 1990 and identify patterns of spread among countries of endemicity. Transitions between different host species are also enumerated. Outbreaks in North Africa appear to have origins in countries immediately south of the Sahara, whereas those in the Middle East are more often from East Africa. The results of the analysis of spread within sub-Saharan Africa are consistent with it being driven by relatively short-distance movements of animals across national borders, and the analysis of host species transitions supports the role of the African buffalo (*Syncerus caffer*) as an important natural reservoir.

IMPORTANCE Foot-and-mouth disease virus is a livestock pathogen of major economic importance, with seven distinct serotypes occurring globally. The SAT 2 serotype, endemic in sub-Saharan Africa, has caused a number of outbreaks in North Africa and the Middle East during the last decades, including three separate incidents in 2012. A comprehensive analysis of all available RNA sequences for SAT 2 has not been published for some years. In this work, we performed this analysis using all previously published sequences and 49 newly determined examples. We also used phylogenetic methods to infer the source country for all outbreaks occurring outside sub-Saharan Africa since 1990 and to reconstruct the spread of viral lineages between countries where it is endemic and movements between different host species.

Received 29 July 2013 **Accepted** 17 September 2013 **Published** 22 October 2013

Citation Hall MD, Knowles NJ, Wadsworth J, Rambaut A, Woolhouse MEJ. 2013. Reconstructing geographical movements and host species transitions of foot-and-mouth disease virus serotype SAT 2. mBio 4(5):e00591-13. doi:10.1128/mBio.00591-13.

Editor Julian Parkhill, The Sanger Institute

Address correspondence to Matthew Hall, m.d.hall@sms.ed.ac.uk.

Foot-and-mouth disease (FMD) is a highly contagious disease of cloven-hoofed mammals caused by FMD virus (FMDV), a single-stranded RNA virus of the family *Picornaviridae*. Seven serotypes exist, two of which (O and A) are found almost worldwide. Another, type C, is more geographically restricted and has not been detected anywhere in the world since 2004, while the Asia-1 serotype is normally confined to southern Asia (1, 2). The remaining three serotypes are the three Southern African Territories (SAT) viruses, designated SAT 1, SAT 2, and SAT 3, the first two of which are endemic in countries of Africa south of the Sahara; outbreaks due to SAT 3 in domesticated livestock have been confined to a few countries in southern Africa. SAT 2, the focus of this study, is the SAT serotype most often recorded in domestic animals (3) and is widely distributed across the continent, having been identified as far west as Senegal, east as Ethiopia, and south as South Africa. It is further subclassified into 14 topotypes, designated I to XIV, defined as having 80% nucleotide identity in the VP1 coding region (4, 5).

SAT 2 has crossed the Sahara and caused outbreaks in North Africa and the Middle East on several occasions in recent years. Middle Eastern outbreaks occurred in North Yemen in 1990 (6) and in Saudi Arabia and Kuwait in 2000 (1). In North Africa, it appeared in Libya in 2003 after an apparent absence from the region for around 50 years (7). In 2012, outbreaks occurred in Egypt, the Palestinian Territories, Libya, and Bahrain (6). While it might be surmised that the occurrence of so many events in a single year might be the result of a single introduction that spread further once established north of the Sahara, Ahmed et al. (6) conducted a genetic study of the viruses involved and found that this did not appear to be the case. Although the bulk of the Egyptian and Palestinian isolates are closely related, those from Libya and Bahrain are of quite distinct lineages. The Bahraini virus is even of a different topotype. Furthermore, one of the samples obtained from Egypt proved to be yet another lineage, distinct from the others collected in the country during the epidemic. For the virus to escape from sub-Saharan Africa four times in 1 year is unprecedented, and it has been suggested that the political changes in the region from 2011 onwards (the "Arab Spring") may have played a role (http://www.bbsrc.ac.uk/news/food-security/2012/120613-f-arab-spring-spread-of-animal-disease.aspx), as people and their animals were displaced by conflict or changing governments created new trading relationships and thus new

TABLE 1 Countries and dates of sampling for available FMDV SAT 2

Country	No. of isolates	Date or date range (yr)
Angola	1	1974
Bahrain	5	2012
Botswana	6	1977–1998
Burundi	2	1986–1991
Cameroon	3	2000–2005
Côte d'Ivoire	1	1990
Democratic Republic of the Congo (or Zaire)	2	1974–1982
Egypt	22	2012
Eritrea	3	1998
Ethiopia	25	1990–2010
The Gambia	2	1979
Ghana	2	1990–1991
Kenya	65	1957–2007
Libya	5	2003–2012
Malawi	1	1975
Mozambique	3	1970–1983
Namibia (or South West Africa)	4	1989–1998
Niger	1	2005
Nigeria	2	1975–2007
North Yemen	1	1990
Palestinian Autonomous Territories	1	2012
Rwanda	4	1996–2004
Saudi Arabia	1	2000
Senegal	5	1975–2009
South Africa	31	1959–2001
Sudan (and South Sudan)[a]	6	1977–2010
Tanzania	2	1975–1986
Togo	1	1990
Uganda	13	1975–2007
Zambia (or Northern Rhodesia)[b]	6	1948–1996
Zimbabwe (or Rhodesia)	24	1972–2003
All sequences	250	1948–2012

[a] All isolates sampled before partition of country in 2011.
[b] Isolate RHO/1/48, whose name suggests an origin in modern-day Zimbabwe, was in fact sampled in Northern Rhodesia, which is modern-day Zambia (see http://www.picornaviridae.com/aphthovirus/fmdv/fmd_history.htm).

pathways for pathogens to follow. For example, Kandeil et al. (8) note that cattle imports to Egypt from other countries in the Nile basin increased following the Egyptian revolution of 2011 due to improved political relationships between the governments involved.

The epidemiology of the SAT serotypes in sub-Saharan Africa is distinct from that for other serotypes in Africa and elsewhere in that there exists a wildlife reservoir in the form of African buffalo (*Syncerus caffer*) in areas where that species is present (9). The disease is very rarely symptomatic in buffalo, and animals can be persistently infected for a period of several years. Since eradication of all infected hosts is therefore not feasible, control has focused on vaccination and prevention of mixing between buffalo and livestock by means of fencing (9, 10). Where SAT serotype epidemics have occurred in locations in proximity to areas with buffalo populations, they have sometimes been linked to compromised fences (11). Since other wild mammals, such as impala (*Aepyceros melampus*) and other antelopes, are susceptible to FMDV, another cause for concern is the ability of these species to jump over fences and spread infection in that way (9).

It has been some time since the last published phylogenetic analyses of all known RNA sequences for SAT 2 (12). Since then, the number of available sequences has almost quadrupled, and information on viruses from a much wider range of locations has been added to nucleotide databases. Reclassification of SAT 2 topotypes has also occurred during that time (4, 5). Recently developed phylogenetic techniques enable analyses such as estimation of change in viral genetic diversity over time (13, 14) and the enumeration of historical changes of discrete character states, such as country of origin or host species, on the phylogenetic tree (15). This study aims to update the complete picture of SAT 2 phylogenetics to include all sequences available in 2013, including some previously unpublished, and to apply the new methods to examine the source of all recorded outbreaks occurring beyond sub-Saharan Africa since 1990, as well as movement patterns of lineages between countries where the virus is endemic and between host species.

RESULTS

The data. All available sequences for the VP1 gene of SAT 2 serotype FMDV were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank/). There were a total of 201 records for distinct isolates available. An additional 49 previously unpublished sequences were also analyzed; information on the origins of these can be found in Table S1 in the supplemental material. This gives a total data set of 250 sequences. All sequences were 648 bp in length, with the exception of five West African examples (all of topotype VI), which were each 651 bp. Table 1 summarizes the data, and Table S2 gives more detailed information. Since all relevant sequences were sampled prior to the partition of Sudan in

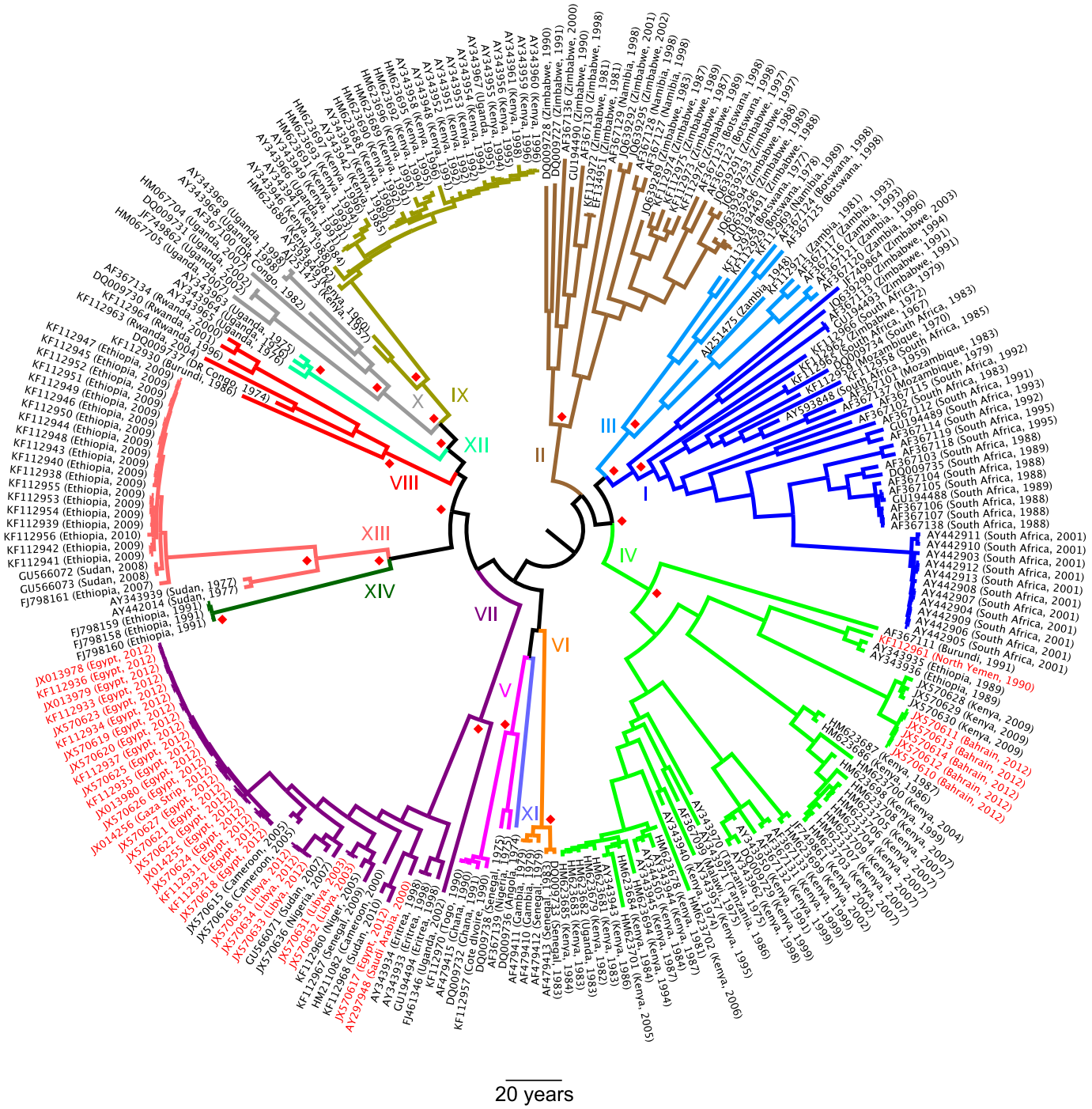


FIG 1 Maximum clade credibility tree of all sequences included in the data set. GenBank accession numbers and countries and dates of sampling are given at the tips; sequences isolated during epidemics in North Africa and the Middle East are in red. Branches are colored and labeled by topotype (I to XIV). Red diamonds indicate clades with a posterior probability of >0.9 (within topotypes, they are omitted for all nodes except for the common ancestor of the topotype).

2011, the country was treated as a single location state for this analysis. Two hundred fifteen sequences were from sub-Saharan countries, and the remaining 35 were from outbreaks in North Africa and the Middle East.

Molecular clock and skyride analysis. A Bayesian phylogenetic analysis was conducted with all 250 VP1 sequences, using the software program BEAST, prerelease version 1.8.0 (16). A relaxed lognormal molecular clock (17) and a Gaussian Markov random field (GMRF) Bayesian skyride tree prior (14) were used. The skyride is a highly parametric method that allows reconstruction of changes in viral population size over the timescale of the tree. Figure 1 is the maximum clade credibility (MCC) tree of this analysis, with branches colored by topotype. The year of the mean time of most recent common ancestor (TMRCA) for all sequences was 1881, with a 95% highest posterior density (HPD) interval from 1853 to 1907.

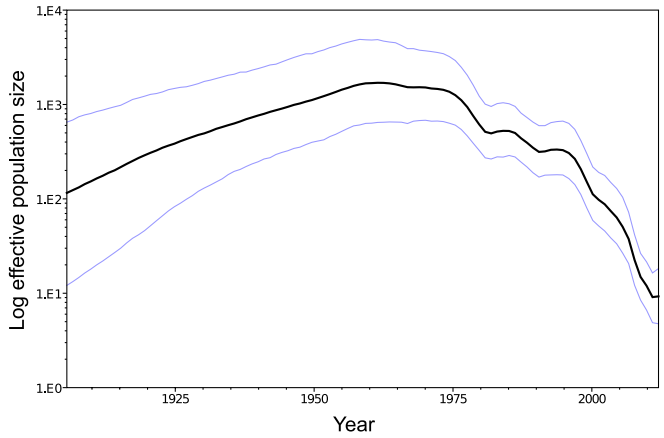


FIG 2 GMRF Bayesian skyride plot of log effective population size against calendar time. Blue lines are the boundaries of the 95% highest posterior density interval.

The estimated parameters of the molecular clock were a mean of 2.45×10^{-3} substitutions per site per year (95% HPD: 1.82×10^{-3}, 3.17×10^{-3}) and a standard deviation of 0.90 (0.72, 1.09). The reconstructed skyride plot can be seen in Fig. 2. Genetic diversity peaked around 1965 and then began to decline, at a rate that increased around 1995.

Phylogeography. In order to explore the origins of the outbreaks outside sub-Saharan Africa, the Monte Carlo Markov chain (MCMC) output from the previous section was used as the set of trees for a discrete-traits phylogeography analysis (15) using the Markov jumps method to reconstruct movements between countries (18). Figure 3 displays the MCC tree, with branches colored by location of sampling for tips and highest posterior probability location for internal nodes. For clarity, sub-Saharan countries have been grouped by United Nations (UN) region.

Figure 4 gives the posterior distributions for the country of origin of each North African and Middle Eastern epidemic occurring since 2000. In a previous analysis of the Egyptian sequences from 2012, Ahmed et al. (6) determined that the isolate EGY/2/2012 (designated strain Alx-12) was most likely the result of a separate introduction to the other sequences from this outbreak (strain Ghb-12). As a result, we examined the origins of these two lineages separately. The possibility that any other epidemic might be the result of multiple introductions was considered, but no such history was reconstructed in any sampled MCMC state. Kenya was overwhelmingly the most likely origin for the 2012 Bahraini virus (posterior probability, 0.89), as was Cameroon for the Ghb-12 lineage of the 2012 Egypt/Palestine outbreak (posterior probability, 0.81). Results were less decisive for the other five outbreaks, with no origin having a posterior probability of more than 0.6. In particular, while the Egyptian Alx-12 lineage appeared most likely to be a descendant of a Sudanese isolate (posterior probability, 0.6), it was also closely related to the virus from 2000 in Saudi Arabia (posterior probability, 0.21). Notably, there was practically no suggestion that any of the 2012 outbreaks were the direct descendants of each other.

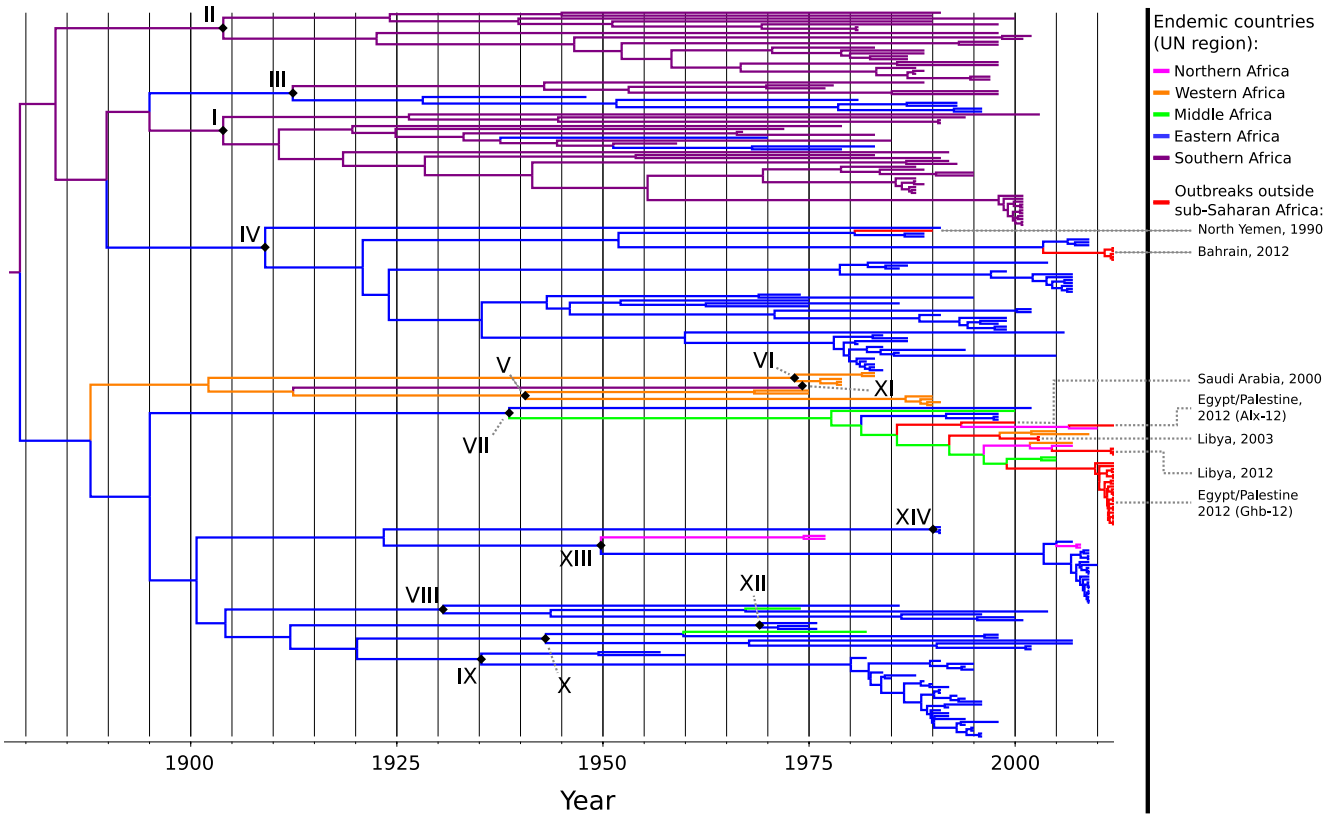


FIG 3 Maximum clade credibility tree of all sequences; branches are colored by UN region within sub-Saharan Africa or red for outbreaks in other areas. Roman numerals and black diamonds indicate nodes representing the common ancestor of each topotype or, where only one sequence was available for that topotype, the tip corresponding to that sequence.

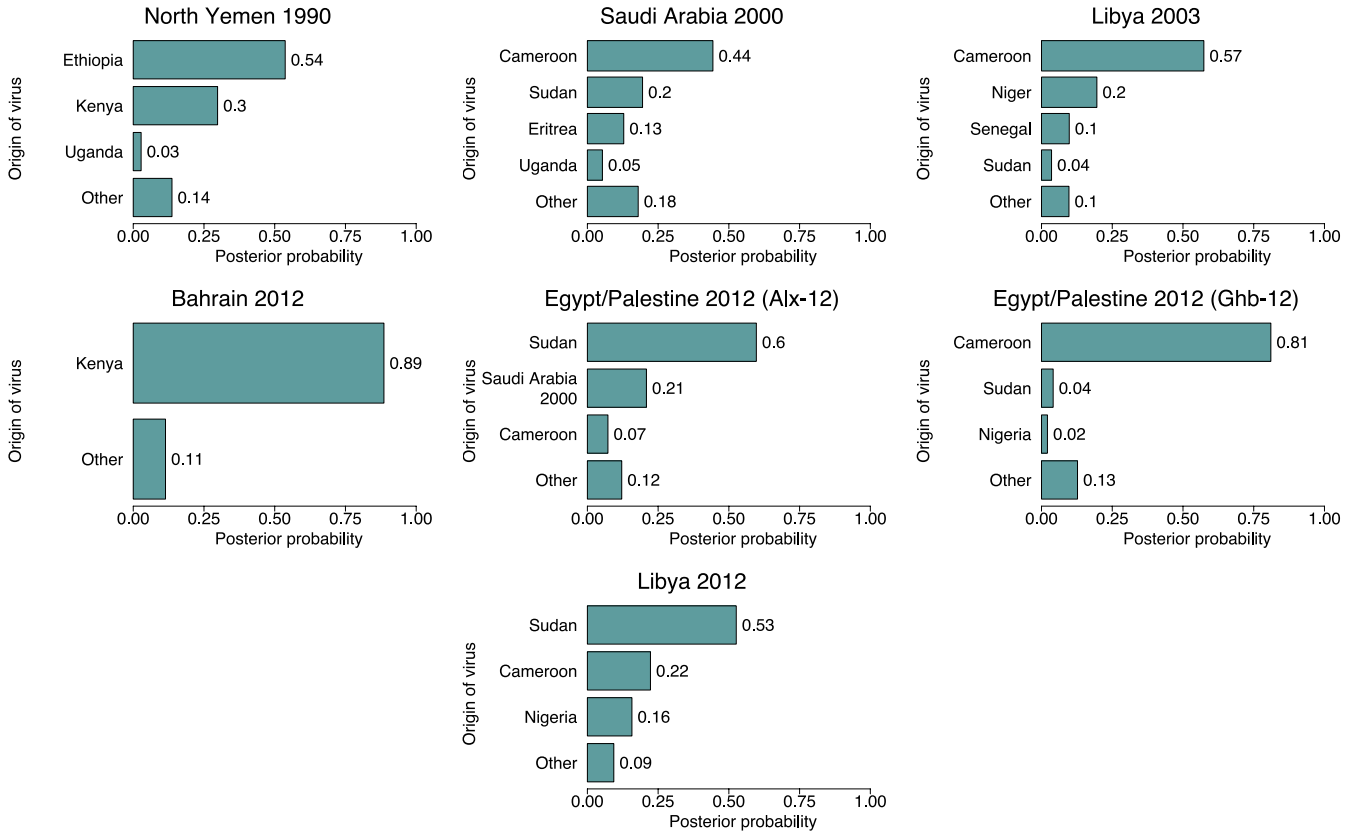


FIG 4 Posterior probability distributions for the countries or epidemic states that were the origins of reconstructed Markov jumps seeding SAT 2 outbreaks in North African and the Middle East, 2000 to 2012. Only origins with a posterior probability of 0.02 or more are shown individually.

A second phylogeography analysis was conducted by restricting the data set to only the 215 sequences from sub-Saharan Africa, in order to identify patterns of movement within the continent. The map in Fig. 5 displays supported nonzero rates of transition (Bayes factor [BF] > 3) between countries of endemicity. Most identified links were across a shared land border; longer-distance links were usually in cases where there were intervening countries from which samples were not available. Longer links also tended to have lower BF support.

Host species analysis. A final discrete-traits analysis was performed to investigate transitions between different host species for the virus. Only 169 sequences had an identified host, which was *S. caffer* in 28 cases, domestic cattle in 130, *A. melampus* in 10, and a pig in 1. The latter was excluded because a single example was unlikely to be adequate for the purpose of investigating the sources of infections in pigs. Figure 6 shows the MCC tree. Branches are colored by host; clades representing topotypes are annotated with a diamond. The most likely root state (the host species of the common ancestor of all known SAT 2 isolates) was *S. caffer*, with a posterior probability of 0.53.

Table 2 summarizes the results of a Markov jumps analysis for changes of host species. The median number of jumps across all trees in the posterior are given for each pair of hosts, along with the posterior probability that the total number of such transitions was nonzero. The median number was nonzero in all cases except transitions from cattle to *A. melampus*, but the only type of transition for which there was 95% support for at least one such jump occurring was from *S. caffer* to cattle.

DISCUSSION

This work has applied recently developed phylogenetic methods to the VP1 gene sequences of all SAT 2 isolates available at the time of writing. It has some limitations, largely imposed by the nature of the available data. The sampling is effectively opportunistic and markedly unbalanced, and the exact effect of this on the discrete-trait inference methods used here for both geography and host species is unclear and warrants investigation in its own right. This makes the results of the host species analysis in particular somewhat incomplete, first because very few countries have available sequences from both cattle and wild animals and second because no sequences at all are available from sheep or goats, despite the hypothesis that they play an important role in the maintenance of FMDV populations (19). In addition, use of simply the country of origin as a location state gives coarse resolution; a lack of links between locations may be simply the result of a lack of sampling in areas sufficiently close to the relevant borders, but restricting to only those sequences where more-detailed location information is available would have greatly decreased the size of the data set.

The VP1 segment was used simply because it has been the most commonly sequenced section of the genome, but use of a larger part would be more suitable and is now more viable in the era of next-generation sequencing. At present, there are eight available sequences for the full SAT 2 genome, and an additional seven for the full coding region (polyprotein gene). While recombination within the structural protein region (VP1 to VP3) appears to be rare, and thus it should not be a cause of concern in interpreting

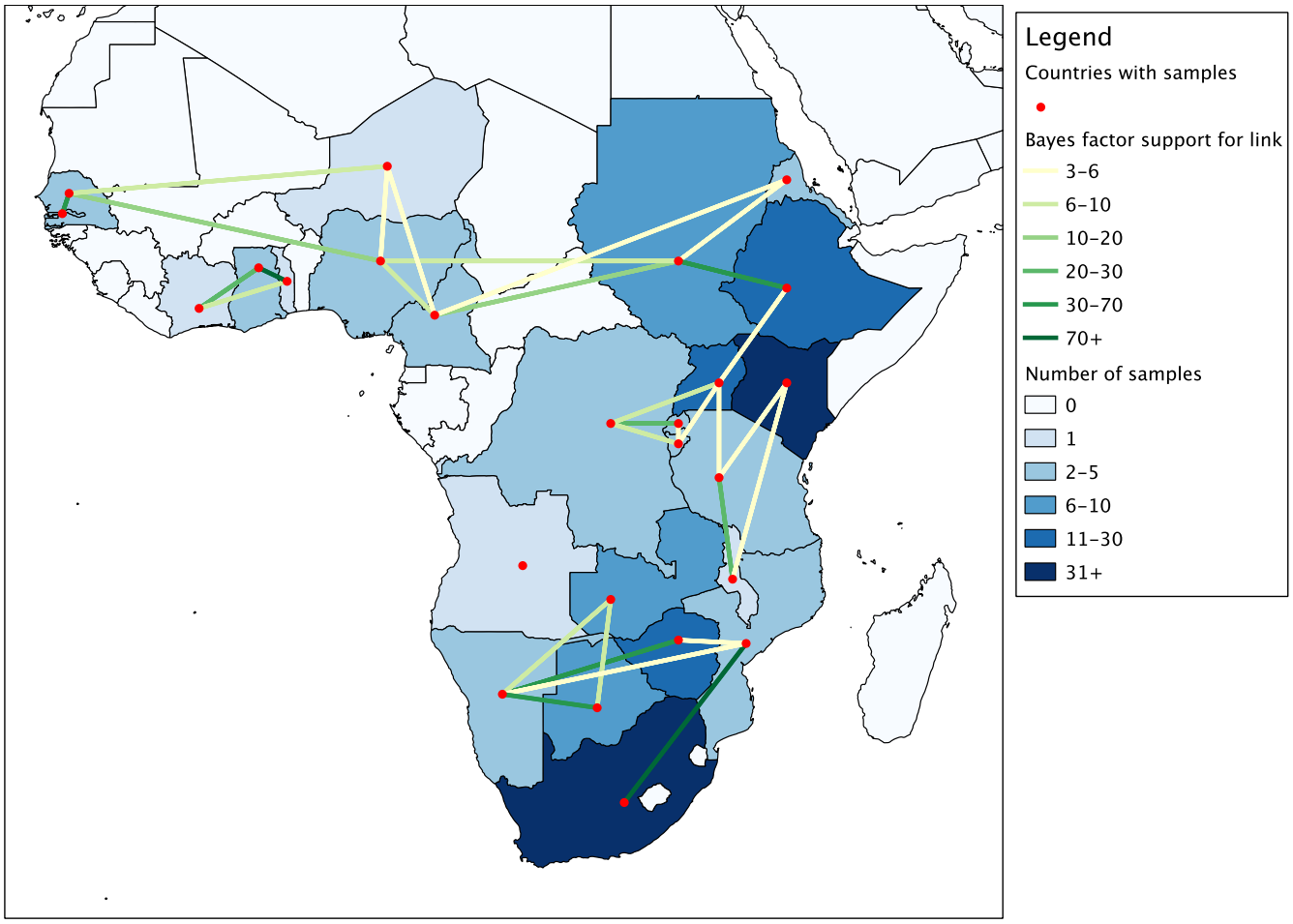


FIG 5 Map of Africa demonstrating links between countries with Bayes factor (BF) support of >3 identified from the BSSVS analysis. Countries are colored by number of sequences available from that location; links are colored by BF value.

this analysis, it is widespread in other parts of the genome (20, 21). This likely renders a naive whole-genome phylogenetic approach inadvisable. Indeed, van Rensburg et al. (22) found that the leader and 3C proteinases of SAT FMDVs displayed branching patterns very different from those of VP1, and it is this recombination that likely explains the findings by Yoon et al. (23) and Lewis-Rogers et al. (24) that when a full-genome analysis is performed, the SAT strains do not form separate clades. However, while the entire genome may not be a good subject for analysis, future work could use the whole structural protein region, rather than just VP1.

The estimated substitution rate of 2.45×10^{-3} substitutions per site per year is very similar to the 2.48×10^{-3} given by Tully et al. (25) for their analysis of the VP1 segment of all FMDV serotypes but considerably higher than their specific estimate for SAT 2 of 1.07×10^{-3}, and the 95% HPD interval of 4.90×10^{-6} to 1.14×10^{-3} given there does not overlap the one found here. That estimate, however, is very imprecise compared to the results for all other serotypes in the same paper, and the data set of 32 sequences used by the authors was also much smaller than ours, covering only 10 of the 14 topotypes. The lower substitution rate estimate in that paper naturally corresponded to an earlier estimated TMRCA of the year 1777, with a 95% HPD interval from 1747 to 1913, also very different from the estimate here, although in this case the HPD interval does overlap ours. Yoon et al. (23) also estimated a lower overall substitution rate (1.46×10^{-3}) for all serotypes, but this analysis was of the full genome, ignoring recombination, and a different rate might be expected. This paper also estimated a much earlier TMRCA for SAT 2 in 1615, with the 95% HPD interval from 1324 to 1866, slightly overlapping ours.

The decline in genetic diversity of FMDV in the latter part of the 20th century has previously been noted by Yoon et al. (23), whose analysis of all seven serotypes also identifies a peak in the middle of the century and a faster decline starting around 2000. A similar peak was also identified by Tully et al. (25), although their analysis suggests a subsequent sharp increase in the last years of the century. A potential explanation for the midcentury decline is the vaccination and fencing measures that have been put in place over the past decades in southern Africa in order to prevent the infection of cattle by wild animals (9, 10). The steeper decline observed starting in the mid-1990s may be a sampling artifact due to the inclusion of a disproportionate number of sequences from comparatively well-sampled epidemics with dates from this time period, since the increased number of coalescent events associated with such data might lead to artificially low estimates of the effective population size. Alternatively, it could reflect a genuine decrease in diversity, possibly due to improving farming practices.

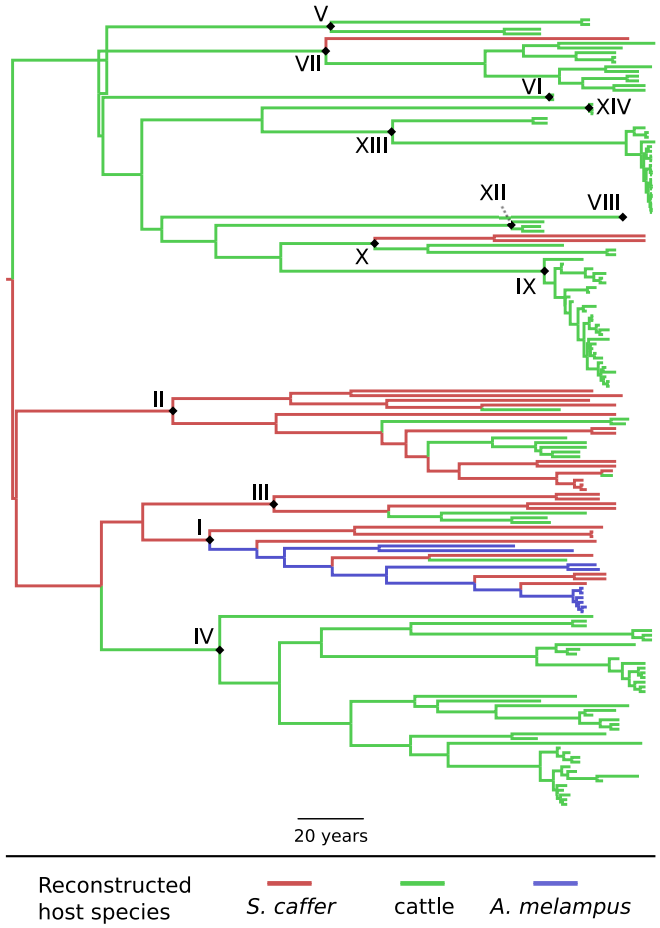


FIG 6 Maximum clade credibility tree of 168 sequences colored by reconstructed host species. Branches are colored by host. Roman numerals and black diamonds indicate nodes representing the common ancestor of each topotype or, where only one sequence was available for that topotype, the tip corresponding to that sequence.

Since FMDV in Africa is presumably generally spread overland by animal movements, the inference of a particular country as the origin of a particular epidemic in this analysis should not be interpreted as it being the last country in which the lineage was present before the start of the epidemic; for example, no strain could have moved directly in this way from Cameroon to Egypt or Libya for the obvious reason that there are intervening countries on any route between them. Instead, this analysis provides a probability distribution of the location of the most-recent ancestor of the outbreak strain that can be identified from the available data; no conclusions can be drawn regarding the route that might have been taken to get from one country to the other. In particular, the wide distribution of topotype VII, from Nigeria to Eritrea, has previously been noted by Bronsvoort et al. (26) and is thought to be the result of extremely long distance cattle movements which are known to occur between Cameroon and Sudan. Thus, although the origins for the Libya 2003 and Ghb-12 outbreaks are suggested to be Cameroon, the lineages could well have first made their way east to Sudan before crossing the Sahara, with Sudan not being identified as their origin because strains more closely related to them than to known Sudanese isolates have never been sampled in that country.

TABLE 2 Median (across all trees in the posterior sample) numbers of reconstructed Markov jumps between each pair of species in the host species analysis[a]

Origin	No. of jumps to destination		
	S. caffer	Cattle	*A. melampus*
S. caffer		10 (1.00)	3 (0.85)
Cattle	5 (0.94)		0 (0.48)
A. melampus	6 (0.88)	1 (0.53)	

[a] This posterior sample of trees is summarized in Fig. 6. Numbers in parentheses are posterior probabilities for at least one such jump having occurred since the time of the common ancestor of the 168 isolates.

Three separate SAT 2 outbreaks in North Africa and the Middle East in a single year, 9 years after the last such recorded event, might seem unlikely to be independent events, but the evidence here adds further weight to the suggestion (6) that these were not the result of a single introduction and that the concurrence is due to coincidence or to regional circumstances that have made such events more likely. If the latter, this situation may not be particular to SAT 2: a new serotype A virus with a probable origin in sub-Saharan Africa was also discovered in Egypt in 2012 (http://www.wrlfmd.org/fmd_genotyping/2012/WRLFMD-2012-00011%20A%20Egypt%202010-2012.pdf), although whether this was a genuinely new introduction in the very recent past seems an open question, given the fairly frequent occurrence of serotype A in the country (27, 28).

The two topotype IV outbreaks, North Yemen 1990 and Bahrain 2012, were determined to have Kenya or (in the former case) Ethiopia as likely origins. The Bahraini isolates came from cattle that had been recently imported from Saudi Arabia (http://www.promedmail.org/direct.php?id=20120507.1125683). It is unlikely that these strains arrived in the Middle East directly from Kenya by sea; Di Nardo et al. (29) describe cattle movement patterns in the region and did not identify such exports. They do, however, identify imports to Yemen and Saudi Arabia from Somalia, a country whose SAT 2 strains have never been sequenced. Type O FMDV outbreaks in Yemen have previously been traced to cattle from eastern Kenya and Ethiopia traded through markets in Somalia (29), so this would seem the most obvious explanation. Identification of which SAT 2 topotypes are in fact present in Somalia would help confirm this. If the 1990 outbreak originated in Ethiopia, then another possible export route would go through Djibouti.

The Alx-12 strain identified in Egypt is genetically distinct from Ghb-12, and the Markov jumps reconstruction suggests that the most likely origin country was Sudan but that it could also be descended from the 2000 Saudi outbreak. Since it is highly unlikely that both Alx-12 and Ghb-12 were the product of a single viral lineage arriving in Egypt, it seems most probable that there was indeed a fourth 2012 viral escape of this serotype from sub-Saharan Africa. While we did identify different most-likely countries of origin for the two strains, this does not rule out the introductions being the result of the import of the same group of infected animals from Sudan, since the Ghb-12 lineage, originating in Cameroon, may have traveled east on its route to Egypt. If there were indeed two separate introduction events, the cause might be the increase in cattle imports to Egypt identified by Kandeil et al. (8).

The close relationship of Alx-12 to the Saudi strain does sug-

gest another possibility, however: that this lineage may have been present but undetected in North Africa and the Middle East since 2000 or even earlier, its detection in 2012 being the result of the increased surveillance connected to the Ghb-12 outbreak. Since other FMDV serotypes are endemic in these areas (1, 30), it is plausible that it was overlooked. In this scenario the virus persisted in the region following the 2000 outbreak or even was present before that. If true, then the virus is likely to have been maintained in sheep or goats, species in which clinical disease is less likely to be apparent (19). Sheep populations have previously been implicated in maintaining FMDV in these areas (30, 31). Further viral samples from the area and from other countries where topotype VII is present would be required to clarify the picture. A question that also arises is why, in this case, the Ghb-12 introduction would cause a rapidly spreading epidemic and disease control emergency while the existing presence of Alx-12 did not.

Aside from the clear difference between Alx-12 and Ghb-12, there was no suggestion that any other outbreak was the result of multiple introductions, and no lineage of the 2012 outbreak was suggested to be the source of any of the others.

It is generally accepted that FMDV is spread locally in Africa by movements of both livestock and wild animals (that it is frequently subclinical in wild *S. caffer* is considered a major challenge to control of the disease [32–35]). The phylogeographical analysis within countries of endemicity presented here lends some formal support to this hypothesis, since movements over large distances were rarely indicated except where there were intervening countries from which no samples were available, and where such links were suggested, the Bayes factor support was usually on the low side (as in the links from Malawi to Kenya and from Mozambique to Namibia). Investigation into whether the long-distance links between Cameroon and Nigeria and more distant countries to both the west and the east are genuine would require sequences from intervening nations, which are currently unavailable for the full VP1 gene. However, as mentioned above, the close relationship between sequences from Cameroon and samples from Eritrea and the 2000 Saudi outbreak were previously noted by Bronsvoort et al. (26), who point out that cattle are indeed traded directly from Sudan to Cameroon and could have carried the virus over this distance. At the time, no sequences from Sudan or the Central African Republic were available, so the authors acknowledged that they were unable to conclusively demonstrate this. The picture remains patchy, but this analysis does include Sudanese sequences, and links from Cameroon to both Eritrea and Sudan are supported, providing some further evidence for this hypothesis.

Because of the geographical distribution of the available sequences, much more information is available for countries in eastern or southern Africa than for western and central areas, where the picture is fragmentary at best. The situation in the countries south of Cameroon is particularly unclear; apart from sequences from the Democratic Republic of the Congo that are most closely related to isolates from its east, the only isolate from this region is a single Angolan example from 1974, the unique available sequence from topotype XI. No strains from Equatorial Guinea, Gabon, or the Congo have ever been sequenced. Whether topotype XI still exists and more generally what the status of SAT 2 is in this region would appear to warrant further investigation.

The situation in West Africa is better; there are in fact around 50 sequences from countries from Cameroon westward for partial sections of the VP1 gene that were ineligible for this analysis due to being insufficiently long. Sangaré et al. (36) performed an initial phylogenetic analysis on most of these; an extension to this analysis could perform the same phylogeographical methods on the shorter sequences from this area only.

As mentioned above, the host species analysis should be interpreted with caution due to the incomplete nature of the sampling. While there is not strong support here for the hypothesis that virus escapes from natural parks in southern Africa are the result of impala jumping fences (34, 35), the only available *A. melampus* sequences are from the Kruger National Park in South Africa and few subsequent cattle sequences are from any country adjacent to the park. While the coloring of branches in Fig. 6 indicates the most probable host for the common ancestor of each topotype, this is unlikely to be reliable, since many topotypes have had isolates sequenced only from cattle, yet there is no reason to believe that they do not also infect buffalo. The role of any other hosts, such as sheep, cannot be investigated. Nevertheless, that SAT 2 originated in *S. caffer* is consistent with the consensus that buffalo are the maintenance host for the SAT strains (33). Subsequent transitions from *S. caffer* to cattle are reconstructed with support at the 95% level for the count being nonzero; this is consistent with previous literature implicating buffalo as the cause of epidemics in southern Africa (11). Transitions from *S. caffer* to *A. melampus* and vice versa and from cattle to *S. caffer* are also frequently reconstructed with considerable posterior support for their occurrence but not reaching the 95% level. That transitions from buffalo to impala, at least, must occur is generally accepted (33, 34). It is also feasible that cattle and impala infect buffalo, but that hypothesis is not necessary to explain the epidemiology of the virus.

In summary, this paper has used up-to-date methods and sequence data to update the picture of the behavior of the SAT 2 serotype on a continental level. Support is given for generally accepted characteristics of the virus: that it is spread over generally short distances by the land movements of infected hosts and that African buffalo are an important maintenance host. The previous consensus that the 2012 outbreak strains are unrelated and probably did not have the same origins has been strengthened by a formal phylogeographical analysis. Evidence is also provided that the decline in FMDV genetic diversity in the latter part of the 20th century applies to this serotype. Future work on this virus would be enabled by further sequencing, perhaps of a larger part of the genome, with a more methodological sampling scheme. This should become more and more feasible as the technology improves.

MATERIALS AND METHODS

The data. Data used were all GenBank records for FMDV serotype SAT 2 that included at least 90% of the VP1 gene (as of May 2013) and sequences for a total of 49 previously unsequenced cell culture-grown type SAT 2 FMDVs that were obtained from the World Reference Laboratory for Foot-and-Mouth Disease Reference Collection. RNA extraction, reverse transcription-PCR (RT-PCR) of the VP1 region, and RNA sequencing of these was performed as previously described (4, 5). Sequences were assembled using SeqMan Pro (Lasergene v.8 package; DNAstar Inc.). GenBank records were examined to exclude duplicates and isolates for which the year of sampling or country of origin were unavailable. Where two or more records from the same isolate were available, the more recently sequenced version was used. Sequences were aligned using the MUSCLE (37) plugin in the software program Geneious 5.6.4 (Biomatters, Ltd.), and trimmed to the VP1 gene only.

Molecular clock and skyride analysis. The BEAST (16) analysis used a GTR+I+G substitution model, a relaxed uncorrelated lognormal molecular clock (17), and a GMRF Bayesian skyride tree prior (14). Multiple Monte Carlo Markov chain (MCMC) runs of 100,000,000 states each and a burn-in of 10% were combined to obtain a set of 9,000 samples with estimated sample sizes of at least 200 for all numerical model parameters. Tracer 1.5 (http://beast.bio.ed.ac.uk/Tracer) was used to reconstruct the skyride plot and investigate parameter estimates.

Phylogeography. The first phylogeographical analysis was performed using the tree set from the previous section as the sample of phylogenies. An asymmetric rate matrix was assumed. Traits were selected depending on the status of the disease in the country of sampling as follows: for samples from areas in sub-Saharan Africa where FMDV is endemic, the country was used. However, sequences taken from epidemics in North Africa and the Middle East were treated as separate traits even where (in the case of Libya) more than one epidemic had occurred in a single country. The Alx-12 and Ghb-12 strains from Egypt in 2012 were also treated as separate traits. This allowed investigation of the source of each epidemic and the two Egyptian lineages separately. Since any given outbreak could not be the origin of an earlier one, the rates of transition between such states in this direction (e.g., from Egypt in 2012 to Libya in 2003) were set *a priori* to be zero. The software program TreeAnnotator 1.7 was used to produce the MCC tree, with branches colored by trait from this analysis.

Geographical movements were reconstructed using the Markov jumps procedure (18) to give times of state changes along each branch of each tree in the posterior output. These were used to estimate a probability distribution for the country of origin of each of the epidemics, as follows: for every tree in the posterior sample, the tips corresponding to all the samples from an epidemic were identified and the node corresponding to their most recent common ancestor found (this was the tip itself in situations where only a single sequence was available for a given epidemic). If the reconstructed location state of this node was not the same as that of the tips, the epidemic was recorded as being the result of multiple introductions in this particular posterior sample. Otherwise, the reconstructed state change that took the lineage into the epidemic state was found, and the trait that was the origin of this jump was recorded. Summarizing this information over all trees from the sample gave the posterior probability distribution of origins.

A second set of phylogenetic trees was produced, using the same molecular clock and tree prior as above, for those sequences coming only from countries of endemicity. A separate phylogeographic analysis was performed on this, using the Bayesian stochastic search variable selection (BSSVS) procedure to identify pairs of countries for which the hypothesis that the rate of movement between them was nonzero was supported by a Bayes factor value greater than 3. For this analysis, a symmetric rate matrix was assumed. The software program Quantum GIS 1.8.0 (http://qgis.osgeo.org) was used to visualize well-supported nonzero rates on a map.

Host species analysis. A final set of trees was produced by further restricting the data set to those sequences from sub-Saharan Africa with an identified host species. Information from GenBank records and the Picornavirus Home Page (http://www.picornaviridae.com/) was used to provide this information. The posterior set of trees from this was used for the host species analysis. Reconstruction of state changes was again performed using Markov jumps, and the number of transitions between each pair of species was counted for all samples from the MCMC and summarized to give the median number of each type of host-to-host transmission taking place over the phylogeny and the posterior probability that at least one event of each type occurred.

Nucleotide sequence accession numbers. The 49 newly determined sequences have been submitted to GenBank with accession numbers KF112928 to KF112976.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00591-13/-/DCSupplemental.

Table S1, PDF file, 0.1 MB.
Table S2, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

M.D.H. was supported by a Ph.D. studentship from the Scottish Government-funded EPIC program. N.J.K. and J.W. were supported by funding from the Department for Environment, Food and Rural Affairs (Defra), United Kingdom (contract no. SE2939 and SE2940).

We thank members of the Rambaut and Leigh Brown groups at Edinburgh for help with the analyses.

REFERENCES

1. **Knowles NJ, Samuel AR.** 2003. Molecular epidemiology of foot-and-mouth disease virus. Virus Res. **91:**65–80.
2. **Sangula AK, Siegismund HR, Belsham GJ, Balinda SN, Masembe C, Muwanika VB.** 2011. Low diversity of foot-and-mouth disease serotype C virus in Kenya: evidence for probable vaccine strain re-introductions in the field. Epidemiol. Infect. **139:**189–196.
3. **Thomson GR, Bastos ADS, Leotta DF, Primozich JF, Beach KW.** 2004. Foot-and-mouth disease, p 1325–1265. *In* Coetzer JAW, Tustin RC, Leotta DF, Primozich JF, Beach KW (ed), Infectious diseases of livestock, vol 2, 2nd ed. Oxford University Press Southern Africa, Oxford, United Kingdom.
4. **Ayelet G, Mahapatra M, Gelaye E, Egziabher BG, Rufeal T, Sahle M, Ferris NP, Wadsworth J, Hutchings GH, Knowles NJ.** 2009. Genetic characterization of foot-and-mouth disease viruses, Ethiopia, 1981–2007. Emerg. Infect. Dis. **15:**1409–1417.
5. **Habiela M, Ferris NP, Hutchings GH, Wadsworth J, Reid SM, Madi M, Ebert K, Sumption KJ, Knowles NJ, King DP, Paton DJ.** 2010. Molecular characterization of foot-and-mouth disease viruses collected from Sudan. Transbound. Emerg. Dis. **57:**305–314.
6. **Ahmed HA, Salem SA, Habashi AR, Arafa AA, Aggour MG, Salem GH, Gaber AS, Selem O, Abdelkader SH, Knowles NJ, Madi M, Valdazo-González B, Wadsworth J, Hutchings GH, Mioulet V, Hammond JM, King DP.** 2012. Emergence of foot-and-mouth disease virus SAT 2 in Egypt during 2012. Transbound. Emerg. Dis. **59:**476–481.
7. **Valarcher JF, Leforban Y, Rweyemamu M, Roeder PL, Gerbier G, Mackay DK, Sumption KJ, Paton DJ, Knowles NJ.** 2008. Incursions of foot-and-mouth disease virus into Europe between 1985 and 2006. Transbound. Emerg. Dis. **55:**14–34.
8. **Kandeil A, El-Shesheny R, Kayali G, Moatasim Y, Bagato O, Darwish M, Gaffar A, Younes A, Farag T, Kutkat MA, Ali MA.** 2013. Characterization of the recent outbreak of foot-and-mouth disease virus serotype SAT2 in Egypt. Arch. Virol. **158:**619–627.
9. **Vosloo W, Bastos AD, Sangare O, Hargreaves SK, Thomson GR.** 2002. Review of the status and control of foot and mouth disease in sub-Saharan Africa. Rev. Sci. Tech. **21:**437–449.
10. **Hunter P.** 1998. Vaccination as a means of control of foot-and-mouth disease in sub-Saharan Africa. Vaccine **16:**261–264.
11. **Vosloo W, Boshoff K, Dwarka R, Bastos A.** 2002. The possible role that buffalo played in the recent outbreaks of foot-and-mouth disease in South Africa. Ann. N. Y. Acad. Sci. **969:**187–190.
12. **Bastos AD, Haydon DT, Sangaré O, Boshoff CI, Edrich JL, Thomson GR.** 2003. The implications of virus diversity within the SAT 2 serotype for control of foot-and-mouth disease in sub-Saharan Africa. J. Gen. Virol. **84:**1595–1606.
13. **Drummond AJ, Rambaut A, Shapiro B, Pybus OG.** 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. **22:**1185–1192.
14. **Minin VN, Bloomquist EW, Suchard MA.** 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. **25:**1459–1471.
15. **Lemey P, Rambaut A, Drummond AJ, Suchard MA.** 2009. Bayesian phylogeography finds its roots. PLOS Comput. Biol. **5:**e1000520. doi:10.1371/journal.pcbi.1000520.
16. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST. Mol. Biol. Evol. **29:**1969–1973.
17. **Drummond AJ, Ho SY, Phillips MJ, Rambaut A.** 2006. Relaxed phylogenetics and dating with confidence. PLOS Biol. **4:**e88. doi:10.1371/journal.pbio.0040088.
18. **Minin VN, Suchard MA.** 2008. Counting labeled transitions in continuous-time Markov models of evolution. J. Math. Biol. **56:**391–412.

19. **Bronsvoort BM, Tanya VN, Kitching RP, Nfon C, Hamman SM, Morgan KL.** 2003. Foot and mouth disease and livestock husbandry practices in the Adamawa Province of Cameroon. Trop. Anim. Health Prod. **35:**491–507.
20. **Carrillo C, Tulman ER, Delhon G, Lu Z, Carreno A, Vagnozzi A, Kutish GF, Rock DL.** 2005. Comparative genomics of foot-and-mouth disease virus. J. Virol. **79:**6487–6504.
21. **Jackson AL, O'Neill H, Maree F, Blignaut B, Carrillo C, Rodriguez L, Haydon DT.** 2007. Mosaic structure of foot-and-mouth disease virus genomes. J. Gen. Virol. **88:**487–492.
22. **van Rensburg HG, Nel LH.** 1999. Characterization of the structural-protein-coding region of SAT 2 type foot-and-mouth disease virus. Virus Genes **19:**229–233.
23. **Yoon SH, Park W, King DP, Kim H.** 2011. Phylogenomics and molecular evolution of foot-and-mouth disease virus. Mol. Cells **31:**413–421.
24. **Lewis-Rogers N, McClellan DA, Crandall KA.** 2008. The evolution of foot-and-mouth disease virus: impacts of recombination and selection. Infect. Genet. Evol. **8:**786–798.
25. **Tully DC, Fares MA.** 2008. The tale of a modern animal plague: tracing the evolutionary history and determining the time-scale for foot and mouth disease virus. Virology **382:**250–256.
26. **Bronsvoort BM, Radford AD, Tanya VN, Kitching RP, Nfon C, Morgan KL.** 2004. Molecular epidemiology of foot-and-mouth disease viruses in the Adamawa province of Cameroon. J. Clin. Microbiol. **42:**2186–2196.
27. **Knowles NJ, Nazem Shirazi MH, Wadsworth J, Swabey KG, Stirling JM, Statham RJ, Li Y, Hutchings GH, Ferris NP, Parlak U, Özyörük F, Sumption KJ, King DP, Paton DJ.** 2009. Recent spread of a new strain (A-Iran-05) of foot-and-mouth disease virus type A in the Middle East. Transbound. Emerg. Dis. **56:**157–169.
28. **Knowles NJ, Wadsworth J, Reid SM, Swabey KG, El-Kholy AA, Abd El-Rahman AO, Soliman HM, Ebert K, Ferris NP, Hutchings GH, Statham RJ, King DP, Paton DJ.** 2007. Foot-and-mouth disease virus serotype A in Egypt. Emerg. Infect. Dis. **13:**1593–1596.
29. **Di Nardo A, Knowles NJ, Paton DJ.** 2011. Combining livestock trade patterns with phylogenetics to help understand the spread of foot and mouth disease in sub-Saharan Africa, the Middle East and South-East Asia. Rev. Sci. Tech. **30:**63–85.
30. **Samuel AR, Knowles NJ, Mackay DK.** 1999. Genetic analysis of type O viruses responsible for epidemics of foot-and-mouth disease in North Africa. Epidemiol. Infect. **122:**529–538.
31. **Rweyemamu M, Roeder P, Mackay D, Sumption K, Brownlie J, Leforban Y, Valarcher JF, Knowles NJ, Saraiva V.** 2008. Epidemiological patterns of foot-and-mouth disease worldwide. Transbound. Emerg. Dis. **55:**57–72.
32. **Ayebazibwe C, Mwiine FN, Tjørnehøj K, Balinda SN, Muwanika VB, Ademun Okurut AR, Belsham GJ, Normann P, Siegismund HR, Alexandersen S.** 2010. The role of African buffalos (*Syncerus caffer*) in the maintenance of foot-and-mouth disease in Uganda. BMC Vet. Res. **6:**54. doi:10.1186/1746-6148-6-54.
33. **Bastos AD, Boshoff CI, Keet DF, Bengis RG, Thomson GR.** 2000. Natural transmission of foot-and-mouth disease virus between African buffalo (*Syncerus caffer*) and impala (*Aepyceros melampus*) in the Kruger National Park, South Africa. Epidemiol. Infect. **124:**591–598.
34. **Vosloo W, Thompson PN, Botha B, Bengis RG, Thomson GR.** 2009. Longitudinal study to investigate the role of impala (*Aepyceros melampus*) in foot-and-mouth disease maintenance in the Kruger National Park, South Africa. Transbound. Emerg. Dis. **56:**18–30.
35. **Thomson GR, Vosloo W, Bastos AD.** 2003. Foot and mouth disease in wildlife. Virus Res. **91:**145–161.
36. **Sangaré O, Bastos AD, Venter EH, Vosloo W.** 2004. A first molecular epidemiological study of SAT-2 type foot-and-mouth disease viruses in West Africa. Epidemiol. Infect. **132:**525–532.
37. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.